LKronbergfinalProject

December 19, 2025

[1]:
```
# Initialize Otter
import otter
grader = otter.Notebook("finalProject.ipynb")
```

1 Final Project: Regression Inference & Classification

Welcome to the Final Project for Data Science for All! This is the final project for our course and with this project you will get to explore a dataset of your choice. By the end of the project, you will have some experience with:

1. Finding a dataset of interest.
2. Performing some exploratory analysis using linear regression and inference.
3. Building a k-nearest-neighbors classifier.
4. Testing a classifier on data.

1.0.1 Logistics

Rules. Don't share your code with anybody. You are welcome to discuss your project with other students, but don't share your project details or copy a project from the internet. This project should be YOUR OWN (code, etc.). If you do base your project on something you learned online or through a generative AI tool such as ChatGPT, make sure to check with your instructor before getting started and reference your sources as part of this project. The experience of solving the problems in this project will prepare you for the final exam (and life). During the final lab session, you will have a chance to share with the whole class.

Support. You are not alone! Come to lab hours, tutoring hours, office hours, and talk to your classmates. If you're ever feeling overwhelmed or don't know how to make progress, we are here to help! Don't hesitate to send an email.

Advice. Develop your project incrementally. To perform a complicated table manipulation, break it up into steps, perform each step on a different line, give a new name to each result, and check that each intermediate result is what you expect. Don't hesitate to add more names/variables or functions if this helps with your analysis or classifier development. Also, please be sure to not re-assign variables throughout the notebook! For example, if you use max_temperature in your answer to one question, do not reassign it later on.

To get started, load `datascience`, `numpy`, and `plots`.

Reading:

- Inference for Regression
- Classification

[2]:
```
# Don't change this cell; just run it. If you need additional libraries for␣
 ↪your project, you can add them to this cell.

import numpy as np
from datascience import *

# These lines do some fancy plotting magic.
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
import warnings
warnings.simplefilter('ignore', FutureWarning)
from matplotlib import patches
from ipywidgets import interact, interactive, fixed
import ipywidgets as widgets
```

2 1. Picking a Dataset

In this project, you are exploring a dataset of your choice. The dataset should be large enough: multiple individuals (rows) with multiple attributes (columns) such that we can try to make a prediction based on the known information in this dataset using linear regression and/or classification. In this first section you will: - find a data set that you are interested in - record the source of where you found it - save it as a .csv file in the same folder as your jupyter notebook. - make sure you can read it in as a table and that your dataset represents a large enough sample for investigating the possible use of regression inference and clasification - Explore the data using visualization techniques learned in this course - Formulate what (which attributes) you would like to investigate using a linear regression model - Formulate what question you would like to answer with a classifier based on this dataset. For example: (1) Is this movie a thriller or a comedy? (2) Is this amazon order Fraudulant or not? (3) Does this patient have cancer or not? See section in the book for more details on Classification - Discuss your choice with you instructor and get approval to get started with section 2

Note 1: If you need guidance on where and how to find a dataset, ask your instructor for help!

Note 2: Your final project conclusion does not necessarily need to show that you have a good regression model or classifier to make predictions! What is important is your own analysis of its potential and limitations when investigating the dataset for making predictions using these techniques

Question 1.1 In the cell below: 1. Read in the dataset you chose as a table 2. Edit the comment to describe where you found this dataset

[3]:
```
# Edit this comment to describe where you found this dataset
# Load your dataset into a table

```

```
my_data_raw = Table().read_table('spotify_streams_2024.csv')
my_data_raw
```

[3]:
```
Track                       | Album Name                                      |
Artist          | Release Date | ISRC         | All Time Rank | Track Score |
Spotify Streams | Spotify Playlist Count | Spotify Playlist Reach | Spotify
Popularity | YouTube Views | YouTube Likes | TikTok Posts | TikTok Likes  |
TikTok Views    | YouTube Playlist Reach | Apple Music Playlist Count | AirPlay
Spins | SiriusXM Spins | Deezer Playlist Count | Deezer Playlist Reach | Amazon
Playlist Count | Pandora Streams | Pandora Track Stations | Soundcloud Streams |
Shazam Counts | TIDAL Popularity | Explicit Track
MILLION DOLLAR BABY          | Million Dollar Baby - Single                    |
Tommy Richman   | 4/26/2024    | QM24S2402528 | 1             | 725.4       |
390,470,936     | 30,716                 | 196,631,588            | 92
| 84,274,754    | 1,713,126     | 5,767,700    | 651,565,900   | 5,332,281,936
| 150,597,040           | 210                        | 40,975        | 684
| 62                    | 17,598,718            | 114                   |
18,004,655      | 22,931                 | 4,818,457          | 2,669,262     |
nan             | 0
Not Like Us                 | Not Like Us                                     |
Kendrick Lamar | 5/4/2024     | USUG12400910 | 2             | 545.9       |
323,703,884     | 28,113                 | 174,597,137            | 92
| 116,347,040   | 3,486,739     | 674,700      | 35,223,547    | 208,339,025
| 156,380,351           | 188                        | 40,778        | 3
| 67                    | 10,422,430            | 111                   |
7,780,028       | 28,444                 | 6,623,075          | 1,118,279     |
nan             | 1
i like the way you kiss me  | I like the way you kiss me                      |
Artemas         | 3/19/2024    | QZJ842400387 | 3             | 538.4       |
601,309,283     | 54,331                 | 211,607,669            | 92
| 122,599,116   | 2,228,730     | 3,025,400    | 275,154,237   | 3,369,120,610
| 373,784,955           | 190                        | 74,333        | 536
| 136                   | 36,321,847            | 172                   |
5,022,621       | 5,639                  | 7,208,651          | 5,285,340     |
nan             | 0
Flowers                     | Flowers - Single                                |
Miley Cyrus     | 1/12/2023    | USSM12209777 | 4             | 444.9       |
2,031,280,633   | 269,802                | 136,569,078            | 85
| 1,096,100,899 | 10,629,796    | 7,189,811    | 1,078,757,968 | 14,603,725,994
| 3,351,188,582          | 394                        | 1,474,799     | 2,182
| 264                   | 24,684,248            | 210                   |
190,260,277     | 203,384                | nan                | 11,822,942    |
nan             | 0
Houdini                     | Houdini                                         |
Eminem          | 5/31/2024    | USUG12403398 | 5             | 423.3       |
107,034,922     | 7,223                  | 151,469,874            | 88
| 77,373,957    | 3,670,188     | 16,400       | nan           | nan
```

```
| 112,763,851            | 182                        | 12,185        | 1
| 82                    | 17,660,624            | 105                    |
4,493,884       | 7,006                 | 207,179            | 457,017       |
nan              | 1
Lovin On Me                | Lovin On Me                                    | Jack
Harlow    | 11/10/2023   | USAT22311371 | 6             | 410.1       |
670,665,438     | 105,892               | 175,421,034            | 83
| 131,148,091   | 1,392,593     | 4,202,367    | 214,943,489   | 2,938,686,633
| 2,867,222,632          | 138                        | 522,042       | 4,654
| 86                    | 17,167,254            | 152                    |
138,529,362     | 50,982                | 9,438,601          | 4,517,131     |
nan              | 1
Beautiful Things           | Beautiful Things                               |
Benson Boone   | 1/18/2024    | USWB12307016 | 7             | 407.2       |
900,158,751     | 73,118                | 201,585,714            | 86
| 308,723,145   | 4,120,760     | nan          | 29,584,940    | 534,915,313
| 4,601,579,812          | 280                        | 383,478       | 429
| 168                   | 48,197,850            | 154                    |
65,447,476      | 57,372                | nan                | 9,990,302     |
nan              | 0
Gata Only                  | Gata Only                                      |
FloyyMenor     | 2/2/2024     | QZL382406049 | 8             | 375.8       |
675,079,153     | 40,094                | 211,236,940            | 92
| 228,382,568   | 1,439,495     | 3,500,000    | 338,546,668   | 3,804,584,163
| 2,112,581,620          | 160                        | 17,221        | 30
| 87                    | 33,245,595            | 53                     |
3,372,428       | 5,762                 | nan                | 6,063,523     |
nan              | 1
Danza Kuduro - Cover       | ŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷ - ŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷŷ - |
MUSIC LAB JPN  | 6/9/2024     | TCJPA2463708 | 9             | 355.7       |
1,653,018,119   | 1                     | 15                     | nan
| nan           | nan           | nan          | nan           | nan
| nan                   | nan                        | nan           | nan
| nan                   | nan                   | nan                    | nan
| nan                    | nan                | nan           | nan
| 1
BAND4BAND (feat. Lil Baby) | BAND4BAND (feat. Lil Baby)                     |
Central Cee    | 5/23/2024    | USSM12404354 | 10            | 330.6       |
90,676,573      | 10,400                | 184,199,419            | 86
| 32,735,244    | 988,682       | 325,800      | 121,574,500   | 974,656,200
| 174,706,874            | 191                        | 3,823         | 117
| 78                    | 10,800,098            | 92                     |
1,005,626       | 842                   | 3,679,709          | 666,302       |
nan              | 1
… (4590 rows omitted)
```

[4]:
```
grader.check("q1_1")
```

[4]:
```
q1_1 results: All test cases passed!
```

Question 1.2: In the following cell, describe each of the variables in the dataset. Are they categorical or numerical? How many observations are there? Add a code cell below this to show how you found the correct number of observations programmatically.

Categorical: Artist, Album Name, Track. Numerical: Spotify Streams, Spotify Popularity, Number of Ratings, Number of Reviews.

Question 1.3: The dependent or response variable of interest is the variable that we will try to classify later in this project. This is the variable that should have two levels that will be used in classification later.

For example: (1) Is this movie a thriller or a comedy? (2) Is this amazon order Fraudulent or not? (3) Does this patient have cancer or not? See section in the book for more details on Classification.

In the code cell below, make any necessary adjustments to your data so that the variable is formatted in this way. Then assign the variable `var` to the column `label` of your table that contains the observations of this variable (note: the label should be a string, so don't forget the quotes!)

[5]:
```
my_data_raw = my_data_raw.drop('TIDAL Popularity', 'Soundcloud Streams',␣
 ↪'SiriusXM Spins',  'AirPlay Spins', 'ISRC', 'Pandora Streams', 'Pandora␣
 ↪Track Stations', 'TikTok Posts', 'YouTube Playlist Reach', 'TikTok Likes',␣
 ↪'YouTube Likes', 'YouTube Views', 'TikTok Views')
my_data_raw
```

[5]:
```
Track                       | Album Name                                    |
Artist          | Release Date | All Time Rank | Track Score | Spotify Streams |
Spotify Playlist Count | Spotify Playlist Reach | Spotify Popularity | Apple
Music Playlist Count | Deezer Playlist Count | Deezer Playlist Reach | Amazon
Playlist Count | Shazam Counts | Explicit Track
MILLION DOLLAR BABY         | Million Dollar Baby - Single                  |
Tommy Richman   | 4/26/2024    | 1             | 725.4       | 390,470,936     |
30,716                  | 196,631,588             | 92                  | 210
| 62                    | 17,598,718            | 114                   |
2,669,262     | 0
Not Like Us                 | Not Like Us                                   |
Kendrick Lamar | 5/4/2024     | 2             | 545.9       | 323,703,884     |
28,113                  | 174,597,137             | 92                  | 188
| 67                    | 10,422,430            | 111                   |
1,118,279     | 1
i like the way you kiss me | I like the way you kiss me                   |
Artemas         | 3/19/2024    | 3             | 538.4       | 601,309,283     |
54,331                  | 211,607,669             | 92                  | 190
| 136                   | 36,321,847            | 172                   |
5,285,340     | 0
Flowers                     | Flowers - Single                              |
Miley Cyrus     | 1/12/2023    | 4             | 444.9       | 2,031,280,633   |
269,802                 | 136,569,078             | 85                  | 394
```

```
| 264                   | 24,684,248            | 210                   |
11,822,942    | 0
Houdini                   | Houdini                                        |
Eminem          | 5/31/2024    | 5             | 423.3       | 107,034,922     |
7,223                   | 151,469,874           | 88                  | 182
| 82                    | 17,660,624            | 105                   |
457,017       | 1
Lovin On Me               | Lovin On Me                                    | Jack
Harlow     | 11/10/2023    | 6             | 410.1       | 670,665,438     |
105,892                 | 175,421,034           | 83                  | 138
| 86                    | 17,167,254            | 152                   |
4,517,131     | 1
Beautiful Things          | Beautiful Things                               |
Benson Boone    | 1/18/2024    | 7             | 407.2       | 900,158,751     |
73,118                  | 201,585,714           | 86                  | 280
| 168                   | 48,197,850            | 154                   |
9,990,302     | 0
Gata Only                 | Gata Only                                      |
FloyyMenor      | 2/2/2024     | 8             | 375.8       | 675,079,153     |
40,094                  | 211,236,940           | 92                  | 160
| 87                    | 33,245,595            | 53                    |
6,063,523     | 1
Danza Kuduro - Cover      | ýýýýýýýýýýýýýýýýýýýýý - ýýýýýýýýýýýýýýýýýý - |
MUSIC LAB JPN   | 6/9/2024     | 9             | 355.7       | 1,653,018,119   |
1                       | 15                    | nan                 | nan
| nan                   | nan                   | nan                   | nan
| 1
BAND4BAND (feat. Lil Baby) | BAND4BAND (feat. Lil Baby)                   |
Central Cee     | 5/23/2024    | 10            | 330.6       | 90,676,573      |
10,400                  | 184,199,419           | 86                  | 191
| 78                    | 10,800,098            | 92                    |
666,302       | 1
… (4590 rows omitted)
```

[6]:
```python
# If you need to make adjustments to your data so that the variable is␣
 ↪formatted in 2 levels add the needed code below, before setting var
import math
def  str_nan_check(a):
    return a != 'nan'

def float_nan_check(a):
    return not(math.isnan(a))

def intify(a):
    return int(a.replace(",", ""))

def floatify(a):
```

```
    return float(a.replace(",", ""))

#$intify(my_data.column("Spotify Popularity").item(0))


my_data = my_data_raw
my_data = my_data.with_columns('All Time Rank', my_data.apply(intify, 'All Time␣
 ↪Rank'))

my_data_raw = my_data_raw.with_column("Top Rank", my_data.column("All Time␣
 ↪Rank")<500)

# my_data = my_data.where('Spotify Playlist Reach', are.below(1e6))
# my_data = my_data.where('Spotify Streams', are.below(1e9))
var = 'Top Rank'
```

[7]:
```
my_data = my_data_raw

my_data = my_data.where('Spotify Popularity', float_nan_check)

my_data = my_data.where('Spotify Streams', str_nan_check)
my_data = my_data.with_columns('Spotify Streams',my_data.apply(intify, 'Spotify␣
 ↪Streams'))

my_data = my_data.where('Amazon Playlist Count', float_nan_check)
my_data = my_data.where('Apple Music Playlist Count', float_nan_check)

my_data = my_data.where('Shazam Counts', str_nan_check)
my_data = my_data.with_columns('Shazam Counts', my_data.apply(intify, 'Shazam␣
 ↪Counts'))

my_data = my_data.where('Spotify Playlist Reach', str_nan_check)
my_data = my_data.with_columns('Spotify Playlist Reach', my_data.apply(intify,␣
 ↪'Spotify Playlist Reach'))

my_data = my_data.where('Spotify Playlist Count', str_nan_check)
my_data = my_data.with_columns('Spotify Playlist Count', my_data.apply(intify,␣
 ↪'Spotify Playlist Count'))

#my_data = my_data.where('Release Date', descending = False)

my_data = my_data.where('Deezer Playlist Count', float_nan_check)
#my_data = my_data.with_columns('Deezer Playlist Count', my_data.apply(intify,␣
 ↪'Deezer Playlist Count'))

my_data_raw = my_data
```

[8]:
```
grader.check("q1_3")
```

[8]:
```
q1_3 results: All test cases passed!
```

Now, we are ready to investigate our data visually!

Question 1.4: Think about the numerical variables in the dataset that might be related to each other. In the cell below, make three different scatter plots that show the relationship between different variables while also displaying how each case is classified.

Use the following line of code:

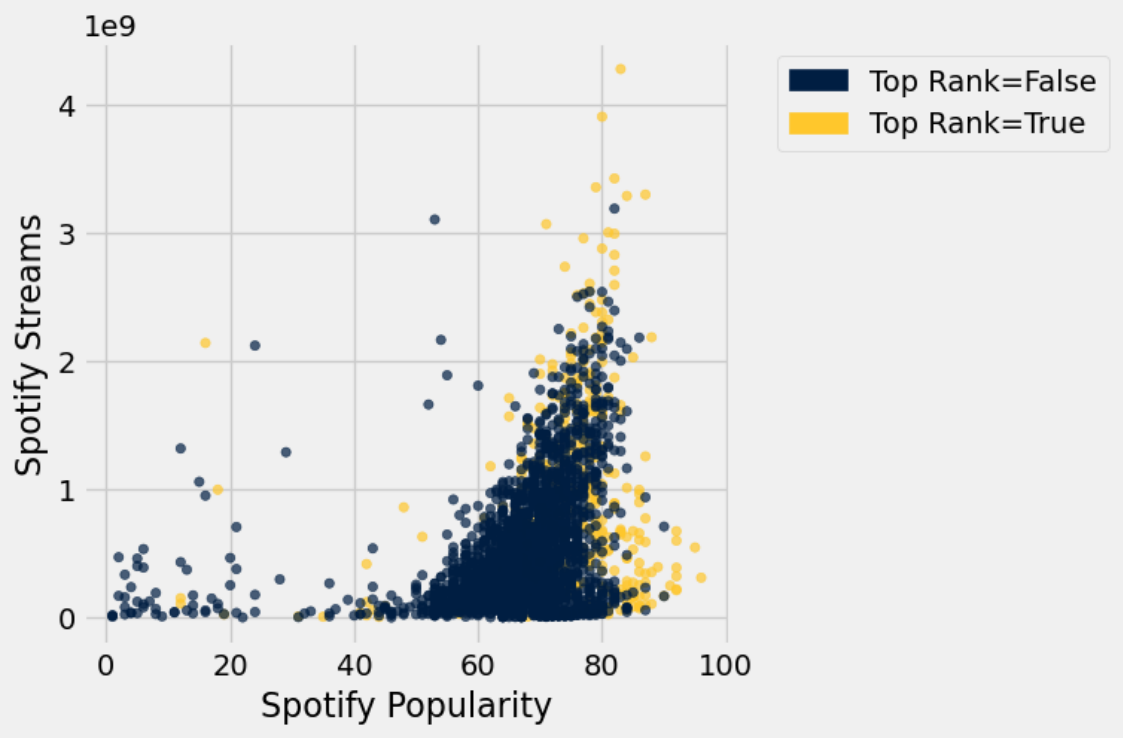
my_data.scatter(`Column 1`, `Column 2`, group=`label`)

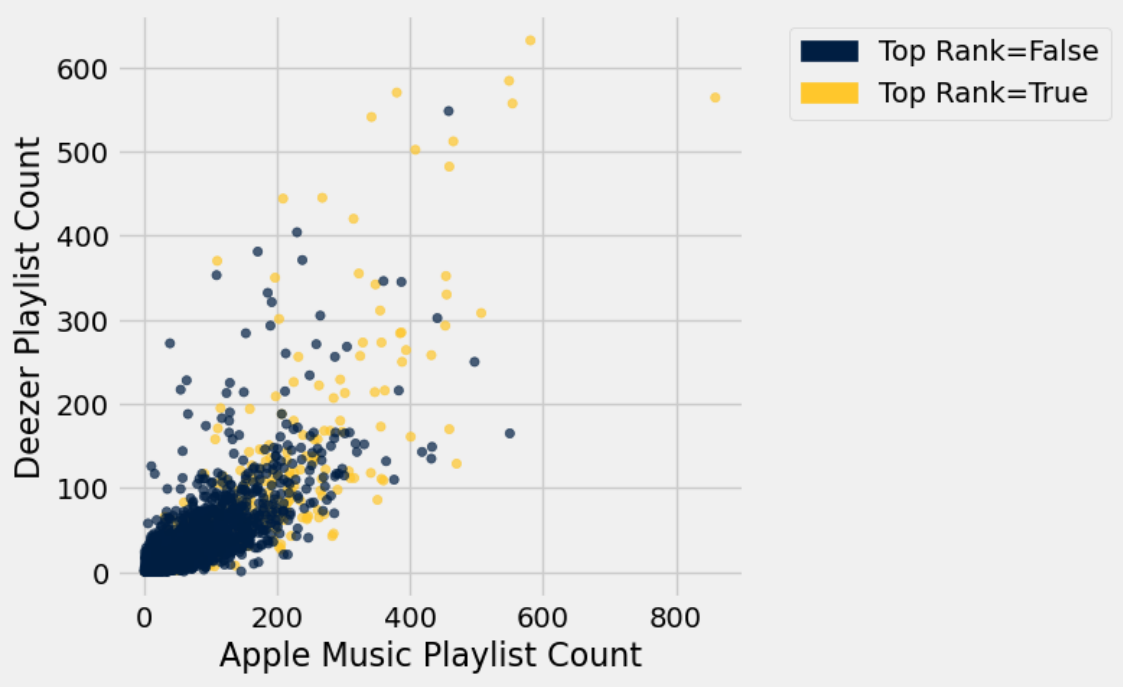
Replace `Column 1` and `Column 2` with the correct column names of numerical variables(features) you would like to investigate. Replace `label` with the column name of the categorical variable you would like to try to classsify.

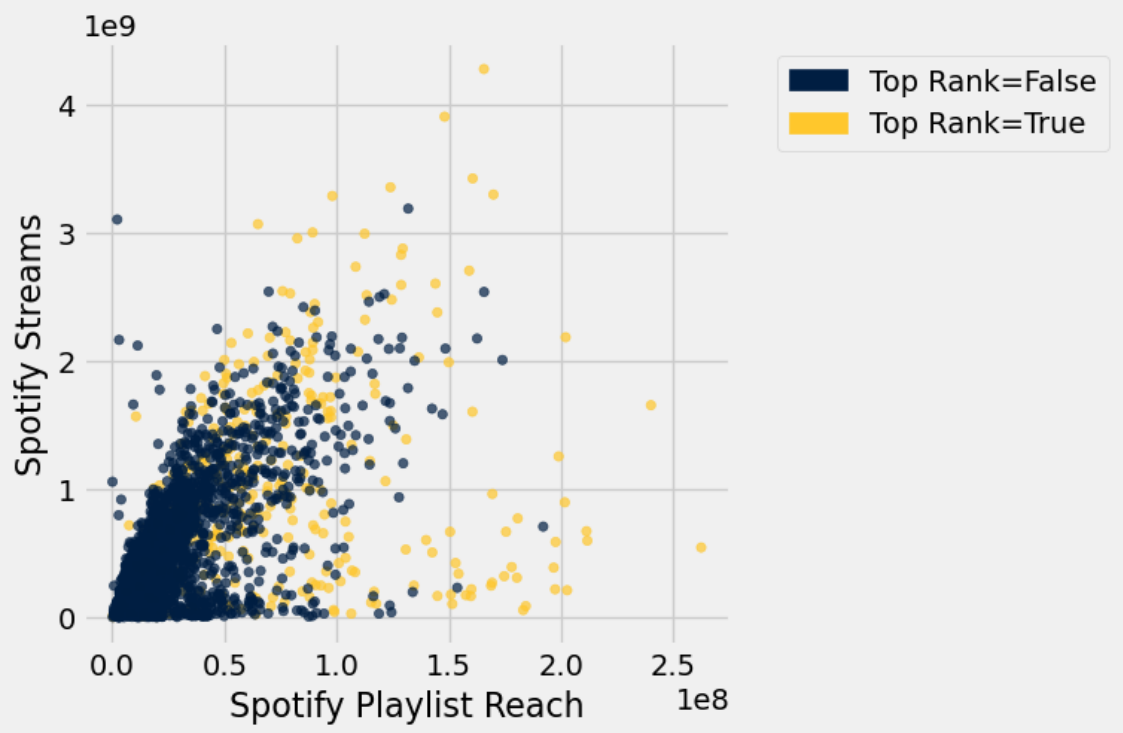
Note: The commented code in the cell below is sample code for this.

[9]:
```
# Here is the code placeholder to use, uncomment the line and adjust according␣
  ↪to the names if the columns in your dataset:
# my_data.scatter("Column 1", "Column 2", group="label")

my_data.scatter("Spotify Popularity", 'Spotify Streams', group = "Top Rank")
my_data_raw.scatter("Apple Music Playlist Count", 'Deezer Playlist Count',␣
  ↪group = "Top Rank")
my_data.scatter("Spotify Playlist Reach", "Spotify Streams", group = "Top Rank")
```




Top Rank=False
Top Rank=True
Deezer Playlist Count
600
500
400
300
200
100
0
0
200
400
600
800
Apple Music Playlist Count


Top Rank=False
Top Rank=True
1e9
Spotify Streams
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0
2.5
1e8
Spotify Playlist Reach

Question 1.5 Describe the three plots from the last question. For each plot, note whether the relationship appears to be linear and whether it is a positive or negative association. Which of the three plots will you look at for linear regression?

I plan to do a linear regression on my Spotify Streams v Spotify Playlist Reach scatter plot relationship.

Question 1.6 In the cell below, formulate the question you would like to try to answer with a classifier that you plan to build.

Type your answer here, replacing this text.

Question 1.7 Set the variable **instructor_signed_off** to 'YES' if you have checked in with your instructor during lab hours.

[10]:
```
instructor_signed_off = 'YES'
```

[11]:
```
grader.check("q1_7")
```

[11]:
```
q1_7 results: All test cases passed!
```

3 2. Regression Inference

To get started, reduce the table with relevant data that you would like to use to evaluate a linear regression model to make a prediction. In this section, you will evaluate this model and set up a hypothesis test to check if there is true correlation/linear association. You will analyze residuals, confidence interval and prediction lines of best fit.

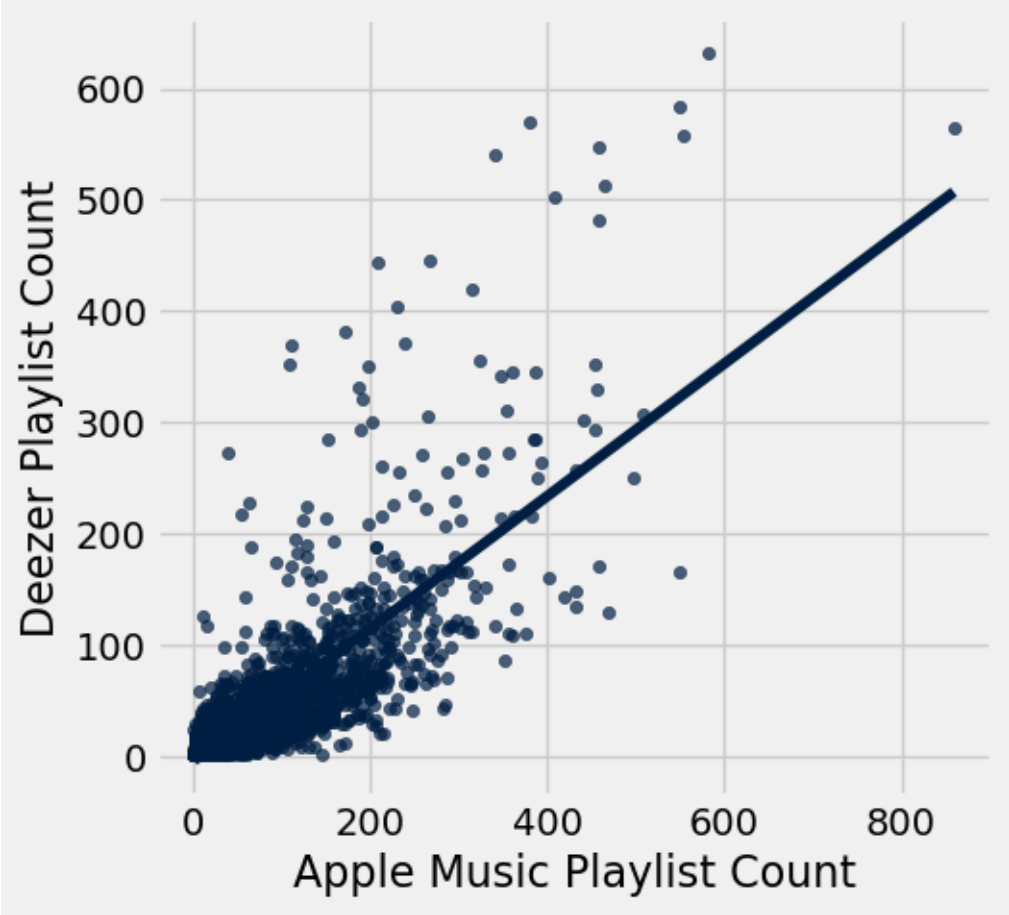
Question 2.1 Copy the cell where you loaded the dataset in section 1, reduce the table to only include the relevant data for what you would like to use in your regression model. The table should only have 2 columns of intest since this is simple linear regression.

[12]:
```
# Copy the cell where you loaded the dataset in section 1 and reduce the table␣
↪to only include the relevant data for linear regression
# You may have to clean your data to get rid of outliers

my_data = my_data.select('Apple Music Playlist Count', 'Deezer Playlist Count')
my_data.show(5)
```

```
<IPython.core.display.HTML object>
```

[13]:
```
# As usual, let's investigate our data visually before analyzing it numerically.
↪
# Just run this cell to plot the relationship between the 2 attribute/columns.
# The scatter plot should look similar to the one you plotted for 1.4.
my_data.scatter(0, 1, fit_line=True)
```



[14]:
```
grader.check("q2_1")
```

[14]:
```
q2_1 results: All test cases passed!
```

Question 2.2:

Use the functions given to assign the correlation between the 2 attributes to the variable `cor`.

The function `correlation` takes in three arguments, a table `tbl` and the labels of the columns you are finding the correlation between, `col1` and `col2`.

[15]:
```
def standard_units(arr):
    return (arr- np.mean(arr)) / np.std(arr)

def correlation(tbl, col1, col2):
    r = np.mean(standard_units(tbl.column(col1)) * standard_units(tbl.
 ↪column(col2)))
    return r
```

```
cor = correlation(my_data, 'Apple Music Playlist Count', 'Deezer Playlist␣
↪Count')
cor
```

[15]:
```
0.7897290078410798
```

[16]:
```
grader.check("q2_2")
```

[16]:
```
q2_2 results: All test cases passed!
```

Can you see a correlation between the 2 variables? If in this sample, we found a linear relation between the two variables, would the same be true for the population? Would it be exactly the same linear relation? Could we predict the response of a new individual who is not in our sample?

Question 2.3: Writing Hypotheses.

Suppose you think the slope of the true line of best fit for the 2 variables is not zero: that is, there is some correlation/association between them. To test this claim, we can run a hypothesis test! Define the null and alternative hypothesis for this test.

Null Hypothesis: There is no correlation between Apple Music Playlist Count and Deezer Playlist Count of songs, what is shown on the graph above is due to random chance. Alternative Hypothesis: There is a positive correlation between Apple Music Playlist Count of songs and Deezer Playlist Count of songs. It is not due to random chance.

Question 2.4:

Maria says that instead of finding the slope for each resample, we can find the correlation instead, and that we will get the same result. Why is she correct? What is the relationship between slope and correlation?

A positive linear slope can mean a positive correlation between factors (as one increases, the other will too). Correlation does not mean causation! Vice versa for a negative linear slope and a negative correlation between compared factors.

Question 2.5: Define the function `one_resample_r` that performs a bootstrap and finds the correlation between the 2 variables in the resample. `one_resample_r` should take three arguments, a table `tbl` and the labels of the columns you are finding the correlation between, `col1` and `col2`.

[17]:
```
def one_resample_r(tbl, col1, col2):
    resample = tbl.sample()
    coR = correlation(resample, col1, col2)
    return coR

# Uncomment the line of code below and change `Column 1` and `Column 2` to␣
↪match your dataset.

one_resample = one_resample_r(my_data, "Apple Music Playlist Count", "Deezer␣
↪Playlist Count")
one_resample
```

[17]:
```
0.76535515519114206
```

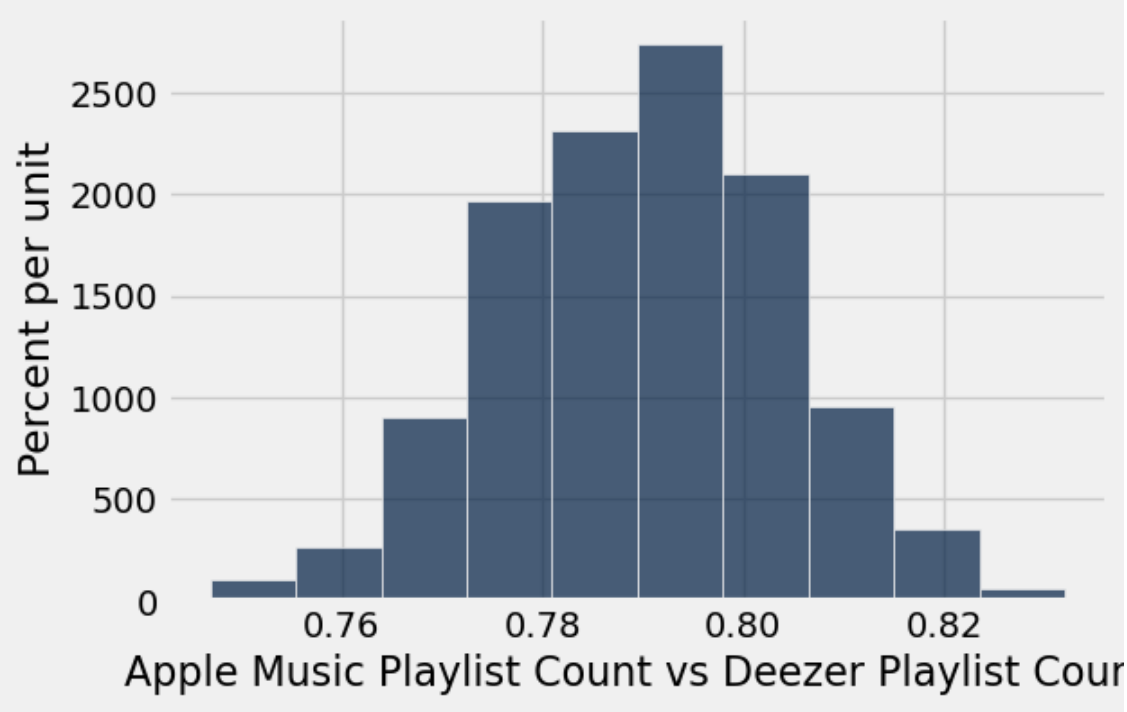
[18]:
```
grader.check("q2_5")
```

[18]:
```
q2_5 results: All test cases passed!
```

Question 2.6: Generate 1000 bootstrapped correlations for the 2 variables, store your results in the array `resampled_correlations`, and plot a histogram of your results.

[19]:
```
resampled_correlations = make_array()
for i in np.arange(1000):
    one_resample = one_resample_r(my_data, "Apple Music Playlist Count",␣
 ↪"Deezer Playlist Count")
    resampled_correlations = np.append(resampled_correlations, one_resample)

# Uncomment the line of code below and change column names to match your dataset
table = Table().with_column("Apple Music Playlist Count vs Deezer Playlist␣
 ↪Count", resampled_correlations)
table.hist()
```



[20]:
```
grader.check("q2_6")
```

[20]:
```
q2_6 results: All test cases passed!
```

Question 2.7: Calculate a 95% confidence interval for the resampled correlations and assign

either `True` or `False` to `reject` if we can reject the null hypothesis or if we cannot reject the null hypothesis using a 5% p-value cutoff.

[21]:
```
lower_bound = percentile(2.5, table.column("Apple Music Playlist Count vs␣
 ↪Deezer Playlist Count"))
upper_bound = percentile(97.5, table.column("Apple Music Playlist Count vs␣
 ↪Deezer Playlist Count"))
reject = False


# Don't change this!
print(f"95% CI: [{lower_bound}, {upper_bound}] , Reject the null: {reject}")
```

```
95% CI: [0.7632038297002663, 0.8167997214704301] , Reject the null: False
```

3.1 Analyzing Residuals

Next, we want to make a prediction for one variable (call this your y variable, or var2) based on the the other (call this your x variable, or var1). First, let's investigate how effective our predictions are.

Question 2.8:

Calculate the slope and intercept for the line of best fit for the 2 variables. Assign these values to `my_slope`, and `my_intercept`respectively. The function `parameters` returns a two-item array containing the slope and intercept of a linear regression line.

Hint 1: Use the `parameters` function with the arguments specified!

*Hint 2: Remember we're predicting the 2nd variable **based off** a first variable. That should tell you what the `colx` and `coly` arguments you should specify when calling `parameters`.*

[22]:
```
# DON'T EDIT THE PARAMETERS FUNCTION
def parameters(tbl, colx, coly):
    x = tbl.column(colx)
    y = tbl.column(coly)

    r = correlation(tbl, colx, coly)

    x_mean = np.mean(x)
    y_mean = np.mean(y)
    x_sd = np.std(x)
    y_sd = np.std(y)

    slope = (y_sd / x_sd) * r
    intercept = y_mean - (slope * x_mean)
    return make_array(slope, intercept)

my_slope = parameters(my_data, "Apple Music Playlist Count", "Deezer Playlist␣
 ↪Count").item(0)
```

```
my_intercept = parameters(my_data, "Apple Music Playlist Count", "Deezer␣
↪Playlist Count").item(1)

my_slope, my_intercept
```

[22]:
```
(0.595446745730829, -4.486499355781241)
```

[23]:
```
real = my_data.column("Deezer Playlist Count")
pred = my_slope * my_data.column("Apple Music Playlist Count") + my_intercept
residuals = real - pred
```

Question 2.9:

Draw a scatter plot of the residuals with the line of best fit for the 2 variables.

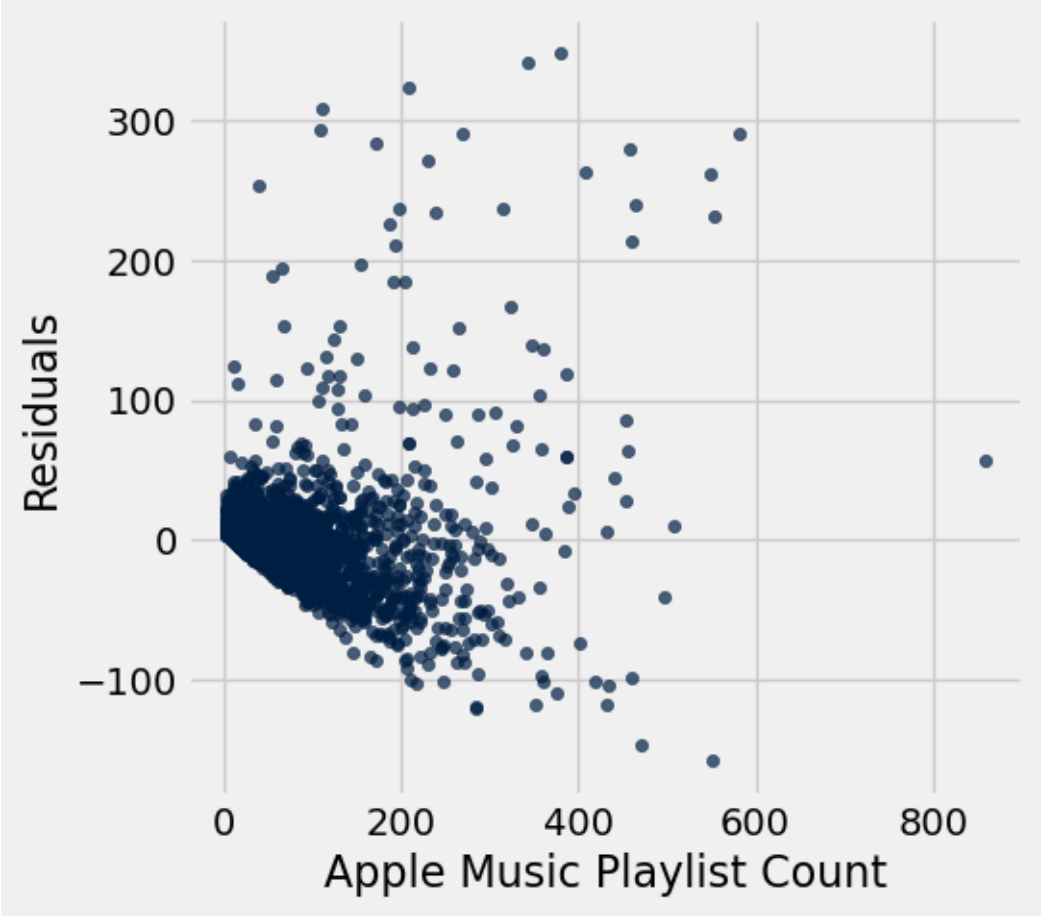
Hint: We want to get the predictions for every data point in the dataset

Hint 2: This question is really involved, try to follow the skeleton code!

[24]:
```
predicted_var2 = my_slope * my_data.column("Apple Music Playlist Count") +␣
↪my_intercept
residuals_var2 = my_data.column("Deezer Playlist Count") - predicted_var2


originalTable_with_residuals = my_data.with_columns("Residuals", residuals_var2)


# Now generate a scatter plot of the residuals!
# Uncomment the line of code below and change "Column 1" to match variable 1␣
↪used in your linear regression analysis
originalTable_with_residuals.scatter("Apple Music Playlist Count", "Residuals")
```



Here's a link to properties of residuals in the textbook that could help out with some questions.

Question 2.10 :

Based on the plot of residuals, do you think linear regression is a good model in this case? Explain.

I think it's deeeeeeeecent. There's a lot of outliers but it's mostly centered around 0 (ergo linear).

Question 2.11 Is the correlation between the residuals and your predictor positive, zero, or negative? Assign `residual_corr` to either 1, 2 or 3 corresponding to whether the correlation between the residuals and your predictor is positive, zero, or negative. Hint: it is ok to check this with Python before answering!

1. Positive
2. Zero
3. Negative

[25]:
```
residual_corr = 1
```

[26]:
```
grader.check("q2_11")
```

[26]:
```
q2_11 results: All test cases passed!
```

3.2 Prediction Intervals

Now, Maria wants to predict the 2nd variable based on a chosen first variable x.

Question 2.12: First, let's identify a value of your choice for x that you want to predict y with and explain in your own words why you chose that value.

x = 200. I chose this x value because it seems reasonable for our graph.

Question 2.13:

Define the function `one_resample_prediction` that generates a bootstrapped sample from the `tbl` argument, calculates the line of best fit for `ycol` vs `xcol` for that resample, and predicts a value based on `xvalue`. Then assign the value you chose for x in Question 2.12 to `chosen_var1`.

Hint: Remember you defined the ***`parameters`*** *function earlier*

[27]:
```
def one_resample_prediction(tbl, colx, coly, xvalue):
    resample = tbl.sample()
    slope = parameters(resample, colx, coly).item(0)
    intercept = parameters(resample, colx, coly).item(1)
    return slope * xvalue + intercept

chosen_var1 = 200

maria_prediction = one_resample_prediction(my_data, "Apple Music Playlist␣
 ↪Count", "Deezer Playlist Count", 200)
maria_prediction
```

[27]:
```
108.90392052414006
```

[28]:
```
grader.check("q2_13")
```

[28]:
```
q2_13 results: All test cases passed!
```

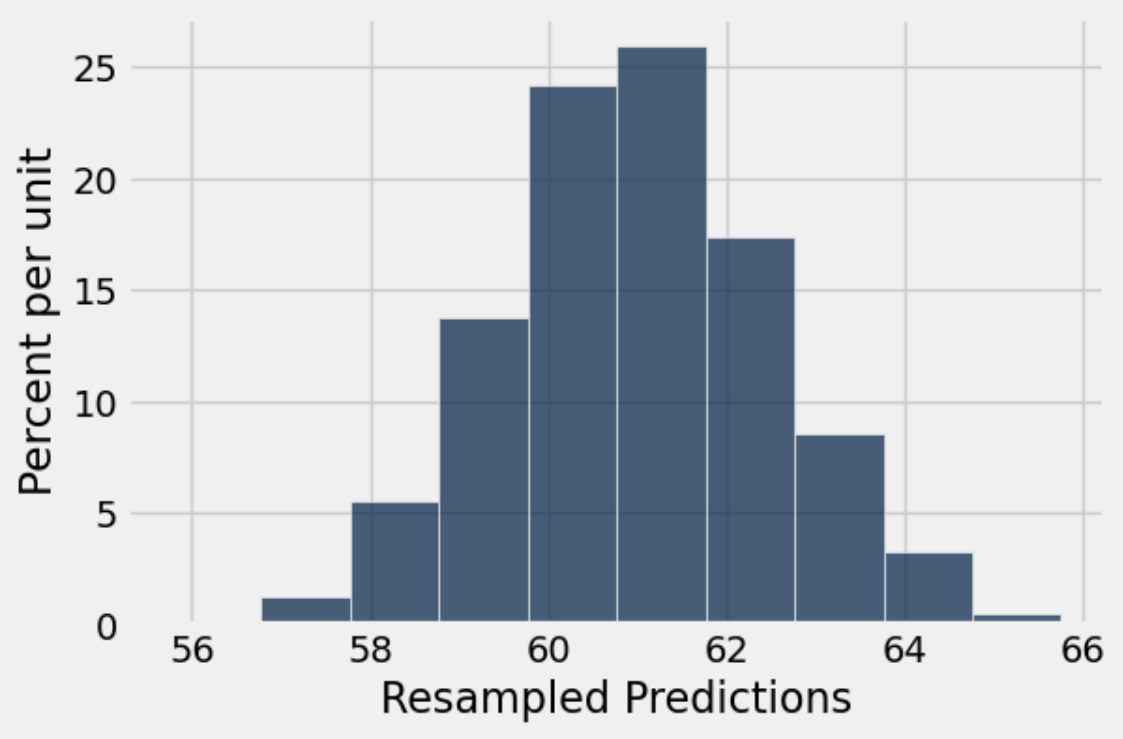
Question 2.14:

Assign `resampled_predictions` to be an array that will contain 1000 resampled predictions for the 2nd variable based on `chose_var1` that you picked, and then generate a histogram of it.

[29]:
```
resampled_predictions = make_array()
for i in np.arange(1000):
    sample = one_resample_prediction(my_data, "Apple Music Playlist Count",␣
 ↪"Deezer Playlist Count", 110)
    resampled_predictions = np.append(resampled_predictions, sample)

# Don't change/delete the code below in this cell, just run to visualize the␣
 ↪distribution
```

```
Table().with_column("Resampled Predictions", resampled_predictions).hist()
```



Question 2.15:

Using `resampled_predictions` from Question 2.14, generate a 99% confidence interval for Maria's prediction.

[30]:
```
lower_bound_maria = percentile(0.5, resampled_predictions)
upper_bound_maria = percentile(99.5, resampled_predictions)


# Don't delete/modify the code below in this cell
print(f"99% CI: [{lower_bound_maria}, {upper_bound_maria}]")
```

```
99% CI: [57.24590718127515, 64.66716850797047]
```

[31]:
```
grader.check("q2_15")
```

[31]:
```
q2_15 results: All test cases passed!
```

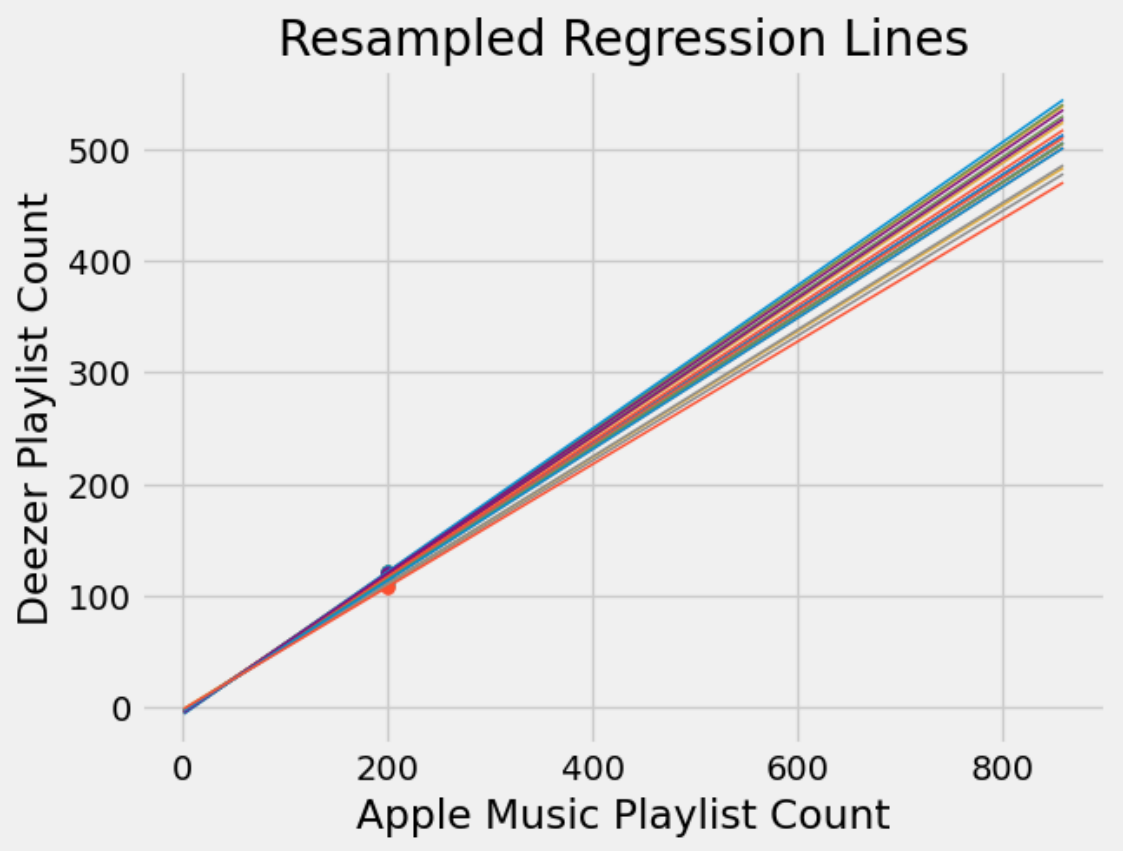
Question 2.16: Uncomment and change the 2 lines of code underneath the TODOs, with the correct `Column 1` and `Column 2`. Then run the following cell to see a few bootstrapped regression lines, and the predictions they make for your chosen value for `chosen_var1` (picked in question 2.13)

[32]:
```
# You don't need to understand all of what it is doing but you should recognize␣
 ↪a lot of the code!
lines = Table(['slope','intercept'])

x=chosen_var1 # This is the value you picked in question 2.14

for i in np.arange(20):
    resamp = originalTable_with_residuals.sample(with_replacement=True)
    # TODO: change Column 1 and Column 2 in the line below and uncomment
    resample_pars = parameters(resamp, "Apple Music Playlist Count", "Deezer␣
 ↪Playlist Count")
    slope = resample_pars.item(0)
    intercept = resample_pars.item(1)
    lines.append([slope, intercept])

lines['prediction at x='+str(x)] = lines.column('slope')*x + lines.
 ↪column('intercept')
# TODO: change Column 1 in the line below and uncomment
xlims = [min(originalTable_with_residuals.column("Apple Music Playlist␣
 ↪Count")), max(originalTable_with_residuals.column("Apple Music Playlist␣
 ↪Count"))]
left = xlims[0]*lines[0] + lines[1]
right = xlims[1]*lines[0] + lines[1]
fit_x = x*lines['slope'] + lines['intercept']
for i in range(20):
    plt.plot(xlims, np.array([left[i], right[i]]), lw=1)
    plt.scatter(x, fit_x[i], s=30)
plt.ylabel("Deezer Playlist Count"); # You can change the label here to be more␣
 ↪descriptive
plt.xlabel("Apple Music Playlist Count"); # You can change the label here to be␣
 ↪more descriptive
plt.title("Resampled Regression Lines");
```



Question 2.17

What are some biases in this dataset that may have affected our analysis? Some questions you can ask yourself are: "is our sample a simple random sample?" or "what kind of data are we using/what variables are we dealing with: are they categorical, numerical, or both (both is something like ordinal data)?".

Hint: you might want to revisit the beginning of this assignment to reread where your data came from and how the table was generated.

The data I'm using from my Streamed Songs dataset is mostly numerical (since I'm mostly dealing with scatterplots and residuals). Some potential sources of bias are where my data came from or how it was collected. Whether or not my data was collected only from a certain region/country, whether or not the people making this dataset liked one artist in particular more than others. How old this dataset is can affect the results.

4 3. Classification

Recommended Reading:

- Classification

This part of the project is about k-Nearest Neighbors classification (kNN), and the purpose is to

reinforce the basics of this method. You will be using the same dataset you picked in section one to complete this part.

We will try to classify our data in 2 classes/groups (labels) based on other variables (features) in our dataset. Go back to question 1.4 and review your answer and your visualization. If it helps copy the code for the visualization below.

4.1 3.1 Splitting the Dataset

Question 3.1. Let's begin implementing the k-Nearest Neighbors algorithm. Define the `distance` function, which takes in two arguments: an array of numerical features (`arr1`), and a different array of numerical features (`arr2`). The function should return the Euclidean distance between the two arrays. Euclidean distance is often referred to as the straight-line distance formula that you may have learned previously.

[33]:
```
def distance(arr1, arr2):
    return np.sqrt(np.sum((arr1 - arr2) ** 2))

# Don't change/delete the code below in this cell
distance_example = distance(make_array(1, 2, 3), make_array(4, 5, 6))
distance_example
```

[33]:
```
5.196152422706632
```

[34]:
```
grader.check("q3_1")
```

[34]:
```
q3_1 results: All test cases passed!
```

4.1.1 Splitting the Dataset

We'll do two different kinds of things with the dataset:

1. We'll build a classifier using the data for which we know the associated label; this will teach it to recognize labels of similar coordinate values. This process is known as *training*.
2. We'll evaluate or *test* the accuracy of the classifier we build on data we haven't seen before.

As discussed in Section 17.2, we want to use separate datasets for training and testing. As such, we split up our one dataset into two.

Question 3.2. Next, let's split our dataset into a training set and a test set. We will start with the full dataset `my_data_raw` (not the one with just the columns used for regression). The table should contain the variable that will be used in classification, it should look like the table after question 1.3.

Now, let's create a training set with the first 75% of the dataset and a test set with the remaining 25% (e.g. if your dataset has 100 rows, 75 rows will be the training set, 25 rows will be the test set). Remember that assignment to each group should be random, so we should shuffle the table first.

Hint: as a first step we can ***shuffle*** *all the rows, then use the* `tbl.take` *function to split up the rows for each table*

[69]:
```
shuffled_table = my_data_raw.sample(with_replacement = False)
train = shuffled_table.take(np.arange(2097))
test = shuffled_table.take(np.arange(2097, my_data_raw.num_rows))

print("Training set:\t",   train.num_rows, "examples")
print("Test set:\t",       test.num_rows, "examples")
train.show(5), test.show(5);
```

```
Training set:    2097 examples
Test set:        699 examples
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

[70]:
```
grader.check("q3_2")
```

[70]:
```
q3_2 results: All test cases passed!
```

4.2 3.2 K-Nearest Neighbors

K-Nearest Neighbors (k-NN) is a classification algorithm. Given some numerical *attributes* (also called *features*) of an unseen example, it decides whether that example belongs to one or the other of two categories based on its similarity to previously seen examples. Predicting the category of an example is called *labeling*, and the predicted category is also called a *label*.

Question 3.3. Assign `chosen_features` to an array of column names (strings) of the features (column labels) from the dataset.

Hint: Which of the column names in the table are the features, and which of the column names correspond to the class we're trying to predict?

Hint: No need to modify any tables, just manually create an array of the feature names!

[71]:
```
chosen_features = make_array('Spotify Streams', 'Spotify Popularity', 'Spotify␣
↪Playlist Count', 'Spotify Playlist Reach')
chosen_features
```

[71]:
```
array(['Spotify Streams', 'Spotify Popularity', 'Spotify Playlist Count',
       'Spotify Playlist Reach'],
      dtype='<U22')
```

Question 3.4. Now define the `classify` function. This function should take in a `test_row` from a table like `test` and classify in using the k-Nearest Neighbors based on the correct `features` and the data in `train`. A refresher on k-Nearest Neighbors can be found here.

Hint 1: The `distance` function we defined earlier takes in arrays as input, so use the `row_to_array` function we defined for you to convert rows to arrays of features.

Hint 2: The skeleton code we provided iterates through each row in the training set.

[72]:
```
def row_to_array(row, features):
    """Converts a row to an array of its features."""
    arr = make_array()
    for feature in features:
        arr = np.append(arr, row.item(feature))
    return arr

def classify(test_row, k, train, features):
    test_row_features_array = row_to_array(test_row, features)
    distances = make_array()

    for train_row in train.rows:
        train_row_features_array = row_to_array(train_row, features)
        train_row_features_array
        row_distance = distance(train_row_features_array,␣
 ↪test_row_features_array)
        distances = np.append(distances, row_distance)
    train_with_distances = train.with_columns("Distances", distances)
    nearest_neighbors = train_with_distances.sort("Distances", descending =␣
 ↪True).take(np.arange(k))
    most_common_label = nearest_neighbors.group('Top Rank').sort('count',␣
 ↪descending = True)
    return most_common_label.column(0).item(0)

# Don't modify/delete the code below
first_test = classify(test.row(0), 5, train, chosen_features)
first_test
```

[72]:
```
True
```

4.2.1 Evaluating your classifier

Now that we have a way to use this classifier, let's focus on the 3 Nearest Neighbors and see how accurate it is on the whole test set.

Question 3.5. Define the function `three_classify` that takes a `row` from `test` as an argument and classifies the row based on using 3-Nearest Neighbors. Use this function to find the `accuracy` of a 3-NN classifier on the `test` set. `accuracy` should be a proportion (not a percentage) of the test data that were correctly predicted.

Hint: You should be using a function you just created!

Note: Usually before using a classifier on a test set, we'd classify first on a "validation" set, which we then can modify our training set again if need be, before actually testing on the test set. You don't need to do that for this question, but please keep this in mind for future courses.

[ ]:
```
def three_classify(row):
    return classify(row, 3, train, chosen_features)
predictions = test.apply(three_classify)
```

```
test_with_prediction = test.with_columns("Predictions", predictions)
labels_correct = test.column("Top Rank")
accuracy = np.count_nonzero(predictions == labels_correct)
accuracy
```

Question 3.6. An important part of evaluating your classifiers is figuring out where they make mistakes. Assign the name `test_correctness` to the test_with_prediction table with an additional column `'Was correct'`. The last column should contain `True` or `False` depending on whether or not our classifier classified correctly. *Note:* You can either include all of the columns from the test_with_prediction table or just the columns representing the features used by the classifier.

[ ]:
```
# Feel free to use multiple lines of code
# but make sure to assign test_correctness to the proper table!
test_correctness = test_with_prediction.with_columns("Was correct",␣
 ↪(predictions == labels_correct))
test_correctness.sort('Was correct', descending = False).show(15)
```

Question 3.7. Do you see a pattern in the rows that your classifier misclassifies? In two sentences or less, describe any patterns you see in the results or any other interesting findings from the table above.

I notice that my predictions are often true and my was correct is often false, so they don't match that much, which corresponds with my low accuracy (26, or 13%).

Question 3.8. Why do we divide our data into a training and test set? What is the point of a test set, and why do we only want to use the test set once? Explain your answer in 3 sentences or less.

Hint: Check out this section in the textbook.

It's to simulate real-world use and give us a more unbiased prediction.

Question 3.9. Why do we use an odd-numbered `k` in k-NN? Explain.

We use an odd numbered k so that we don't get a 50/50 answer. No matter what we will get one that's greater thant the other in our simulation.

At this point, you've gone through one cycle of classifier design. Let's summarize the steps: 1. From available data, select test and training sets. 2. Choose an algorithm you're going to use for classification. 3. Identify some features. 4. Define a classifier function using your features and the training set. 5. Evaluate its performance (the proportion of correct classifications) on the test set.

4.3 4. Explorations

Now that you know how to evaluate a classifier, it's time to build a better one.

Question 4.1:

Develop a classifier with better test-set accuracy than `three_classify`. Your new function should have the same arguments as `three_classify` and return a classification. Name it `another_classifier`. Then, check your accuracy using code from earlier.

You can use more or different features, or you can try different values of `k`. (Of course, you still have to use `train` as your training set!)

Make sure to create new variable names where needed, don't reassign any previously used variables here, such as `accuracy` from the section 3.

[ ]:
```
# Run this cell to remember what your accuracy was in the first attempt.

accuracy
```

[ ]:
```
# Feel free to add or change this array to improve your classifier
# Note that you can either use the original chosen_features or create a new␣
 ↪list below

new_features = make_array("Spotify Playlist Count", "Amazon Playlist Count",␣
 ↪"Deezer Playlist Count", "Apple Music Playlist Count")

def another_classifier(row):
    return classify(row, 3, train, new_features)

new_test_with_prediction = test.apply(another_classifier)
new_labels_correct = test.column("Top Rank")
new_accuracy = np.count_nonzero(new_test_with_prediction == new_labels_correct)
new_accuracy
```

[ ]:
```
# Now that we looked at the accuracy, let's analyze correctness of your new␣
 ↪classifier
# Use this coding cell to explore your data/ this cell will not be graded
# You are free to use as many lines of code as you would like
# you could look at a sorted table again just like in 3.6!
```

Question 4.2

Did your new classifier work better? Do you see a pattern in the mistakes your new classifier makes? What about in the improvement from your first classifier to the second one? Describe in two sentences or less.

Hint: You may not be able to see a pattern.

My new classifier works WAY better! I went from 13% to 84% accuracy.

Question 4.3

Briefly describe what you tried to improve your classifier. Any other ideas on how you could make a better classifier?

I tried using the playlist song count to classify how popular (top ranked) a song is, and it worked pretty well! (84% right!) I can't currently think of anything to make my classifier better, maybe incldluing Spotify Popularity as well? (Tried that just now and no luck).

Question 4.4: Misclassification and errors in classifiers happens all the time and can really affect an individual. When applying machine learning and building classifiers, we all need to do our

best to minimize misclassification and make sure we are transparent about the accuracies of what we built. Have you ever experienced something like this in real life before, where something was classified incorrectly? If not, can you think of an example where misclassification could really affect an individual?

I can't personally think of any machine learning misclassification in my life, but I can think of how errors and inaccuracies can harm an individual. For example: medicine. If machine learning has any bias or any fault in it's programming many people can be misdiagnosed, or maybe their severe health problems could go completely ignored. Accuracy in classification this example is very important.

Question 4.5: We hope you enjoyed the project! You made it to the concluding question.

In a couple of sentences, share what you learned about your dataset while exploring its potential for prediction and classification.

I learned how to make classifiers for a dataset and then make that classifier more accurate with different features and k values. I learned what a features and k values are! Overall a very fun (and a little stressful) projecta and class.

Congratulations: You're DONE with the final project notebook! Nice work. Time to submit.

[ ]:

[ ]: