CSC 4780/6780
Fall 2022
Homework 2

Priyanka Mahendra Ghare
pghare1@student.gsu.edu

September 7, 2022

This homework is due at 11:59 pm on Tuesday, Sept 6. It must be uploaded to iCollege by then. No credit will be given for late submissions. A solution will be released by noon on Wednesday, Sept 7.

Once again: it is a good idea to get this done and turn it in early.

1 Purpose

One of the first things you will do when exploring a new dataset is make some graphs that will give you some intuitive feel for what the data contains.

Also, the last thing you typically do on a project is make the data visualizations that will help your clients understand and believe your analysis.

You will also calculate a gradient. We will be using gradient descent a lot. The reasons will make a lot more sense if you understand gradients.

2 Study

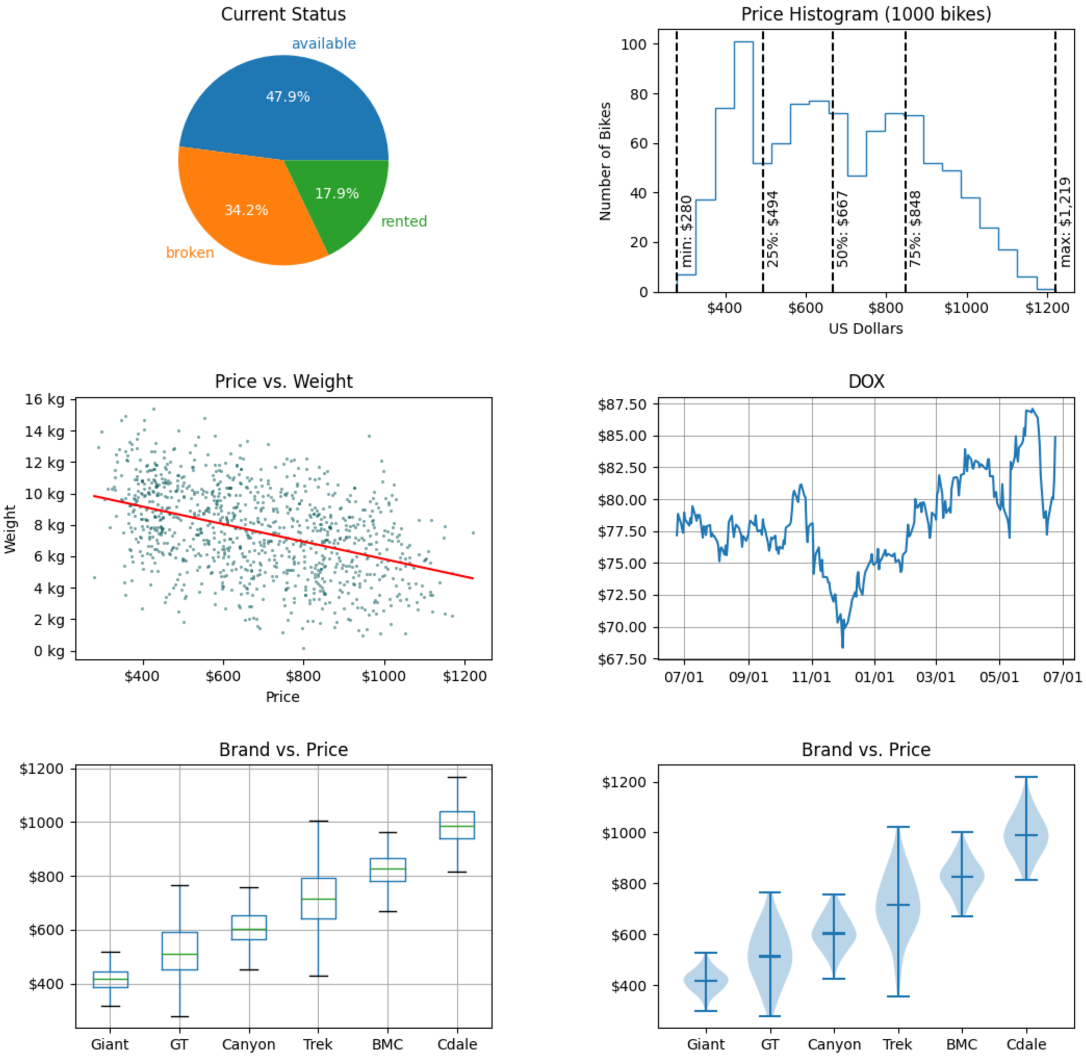
Read pages 133 - 203 in *Practical Data Science with Python.*

3 Make plots with Matplotlib

9 points Create a python program called `make_plots.py` that does the following:

- read `bikes.csv` and `DOX.csv` into pandas dataframes
- divide a matplotlib figure into 3 rows and 2 columns of subplots
- make a pie chart of the statuses of the bikes
- make a histogram of the prices of the bikes
- make a scatter plot of the price vs. the weight of each bike
- make a time-series plot of the price of the DOX stock price
- make a box plot showing the range of prices for each brand
- make a violin plot showing the range of prices for each brand
- save the entire figure into a file called `plots.png`

Matplotlib has lots of options, and an important goal of this is to get you to explore some of those options. Try to make `plots.png` look like this:



Your score will be based upon how well your code generates a plot that matches this.

Your code should assume nothing about the data except the names of the columns. That is, don't hard code any other assumptions about the data in your program.

We have not covered linear regression yet, so here is some code you can use. Assuming that you have loaded the `bikes.csv` file into a pandas dataframe called `df`, this code will give you the slope and y-intercept of the red line in the scatter plot:

```
from sklearn.linear_model import LinearRegression

df = ...
```

```
# Get data as numpy arrays
X =  df['purchase_price'].values.reshape(-1, 1)
y = df['weight'].values.reshape(-1, 1)

# Do linear regression
reg = LinearRegression()
reg.fit(X, y)

# Get the parameters
slope = reg.coef_[0]
intercept = reg.intercept_
print(f"Slope: {slope}, Intercept: {intercept}")
```

Do this work by yourself. Stackoverflow is OK. A hint from another student is OK. Looking at another student's code is *not* OK.

My solution is less than 80 lines of code.

4 Derive a gradient

1 point

Let $f : R^3 \to R$ be given by

$$f(x, y, z) = y \sin(5x) + e^{yz} + \ln z$$

What is the gradient?

Answer: $\nabla f = \left(\frac{\partial i}{\partial x} + \frac{\partial j}{\partial y} + \frac{\partial k}{\partial z} \right) f(x, y, z)$
$\frac{\partial f}{\partial x} = 5y \cos 5x$
$\frac{\partial f}{\partial y} = \sin 5x + z \cdot e^{yz}$
$\frac{\theta f}{\partial z} = y \cdot e^{yz} + \frac{1}{z}$
$\therefore \nabla f(x, y, z) = 5y \cos 5x i + (\sin 5x + z \cdot e^y z) j + \left(y \cdot e^y z + \frac{1}{z} \right) k$

(Feel free to use sympy if your calculus is a little rusty.)

Add the solution here in the LaTeX document and build a pdf from it.

5 What to turn in

If your name is Fred Jones, you will turn in a zip file called HW02_Jones_Fred.zip of a directory called HW02_Jones_Fred. It will contain:

- bikes.csv
- DOX.csv
- make_plots.py
- plots.png
- Assignment.pdf

Be sure to format your python code with black before you submit it.

We will unzip the directory and run your code like this:

```
cd HW02_Jones_Fred
python3 make_plots.py
```

6 Criteria for success

And then we will look at the generated plots.png. If your code doesn't run, you will lose points. If plots.png doesn't look basically like target.png, you will lose points.

For the gradient, the vector should be the correct length. Each component should be correct.

7 Extra help

Here is a good video tutorial on Matplotlib: https://youtu.be/UO98lJQ3QGI

Want to get ahead? Web scraping is next: https://youtu.be/tb8gHvYlCFs