COMP 4304 / 6934 – Assignment 6 (7%)

Due: 11:59pm, Mar. 21, 2024

Learning Objectives

The goal of this assignment is to become familiar with interactive visualizations and ipywidgets.

Data

Download the `diamonds.csv`, `olympic_athletes.csv` and `used_cars.csv` data sets from Brightspace.

The `diamonds.csv` data set contains information about diamond sales. Diamonds have a colour, ranging from D (colourless) to J (yellow-ish). Clarity is a measure of any imperfections in the diamond, which go in order from IF (internally flawless), VVS1/2 (very very slightly included), VS1/2 (very slightly included) to S1/2 (slightly included). Carat is a measure of the weight of a diamond. Diamonds are cut to a particular shape and the quality of the cut is measured. Some diamonds are naturally occurring, but some are manufactured.

The `olympic_athletes.csv` data set contains information on the athletes that competed in the Olympics starting from 1896 to 2016. Each row is an athlete, and contains the athlete's name, country, age, height, weight, which Olympics, event and sport they competed in, and medals they won, if any. Note that athletes that compete in multiple events will appear multiple times (once per event).

The `used_cars.csv` data set contains information on the sales of used cars. It includes the price it sold for, the brand of car, its specific make, as well as other details such as its transmission type.

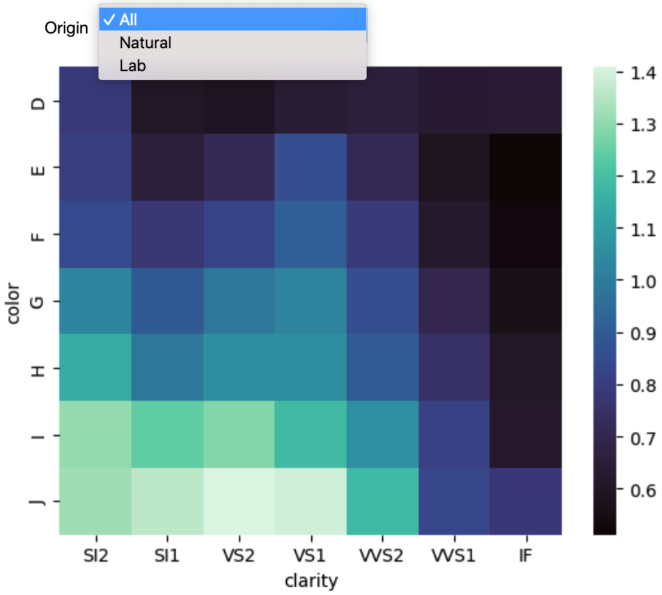
Instructions

Using the provided data sets, create a Jupyter notebook to answer the following questions.

You may only import the pandas, Matplotlib, Seaborn, ipywidgets, and math or NumPy packages.

Question 1: (30 pts)

The following heatmap shows the average carat weight of diamonds relative to their clarity and colour. An interactive dropdown menu allows for the data to be filtered around the origin of the diamonds. The “Natural” option filters to only naturally occurring diamonds, “Lab” filters to manufactured diamonds, and “All’ applies no filtering. In the example below, we can see that the heaviest diamonds are those with the poorest quality.

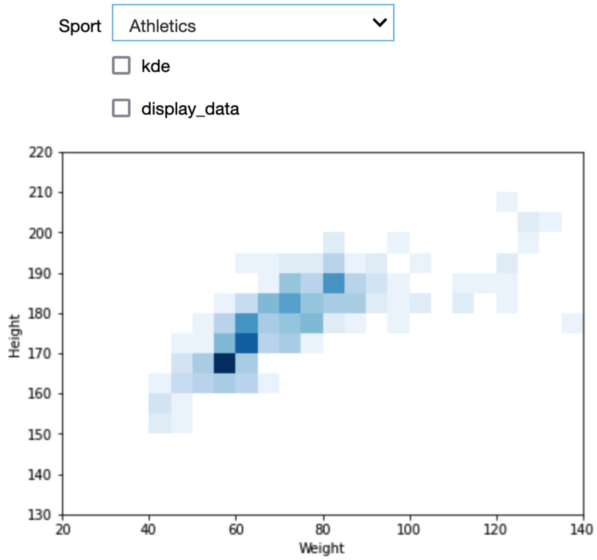


Re-create the above heatmap and interactive control. The heatmap should update based upon the item selected in the dropdown menu.

The colour map is “mako”. The ordering of the colour and clarity should be in ascending order, such that colourless, internally flawless diamonds are in the top right.

Question 2: (40 pts)

The interactive visualization below shows the height and weight distributions for gold-medal winning athletes in the 2000 to 2016 Olympics (inclusive). The default view is a heat map where the data has been binned in value increments of 5, that is, 20 to 25, 25 to 30, 30 to 35, etc. (Hint: Seaborn's heatmap is the hard way to do this. What else can bin data?)



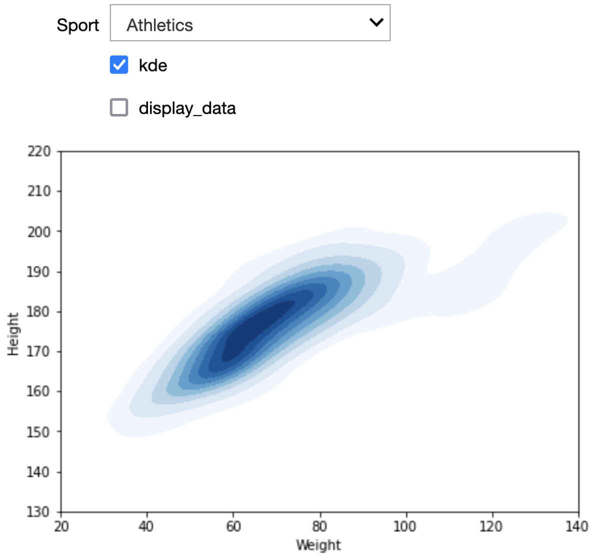
Sport: Athletics

☐ kde

☐ display_data

The dropdown menu filters the athletes to a specific sport category. The options are Swimming, Wrestling, Fencing, Athletics, Shooting, Cycling, Rowing, and Gymnastics, plus a "Combined" option that is the combination of athletes from all of those sports.

The "kde" checkbox will change the visualization to a contour plot of the distribution estimated using kernel density estimation, as shown below.

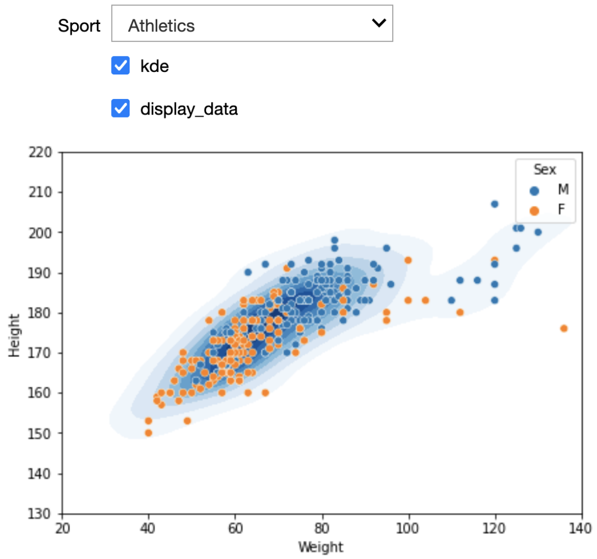


Sport: Athletics

☑ kde

☐ display_data

The “display_data” checkbox will overlay the heat map or contour plot with the raw data points for each individual athlete. These data are colour coded according to the gender of the athlete.



Sport Athletics

☑ kde

☑ display_data

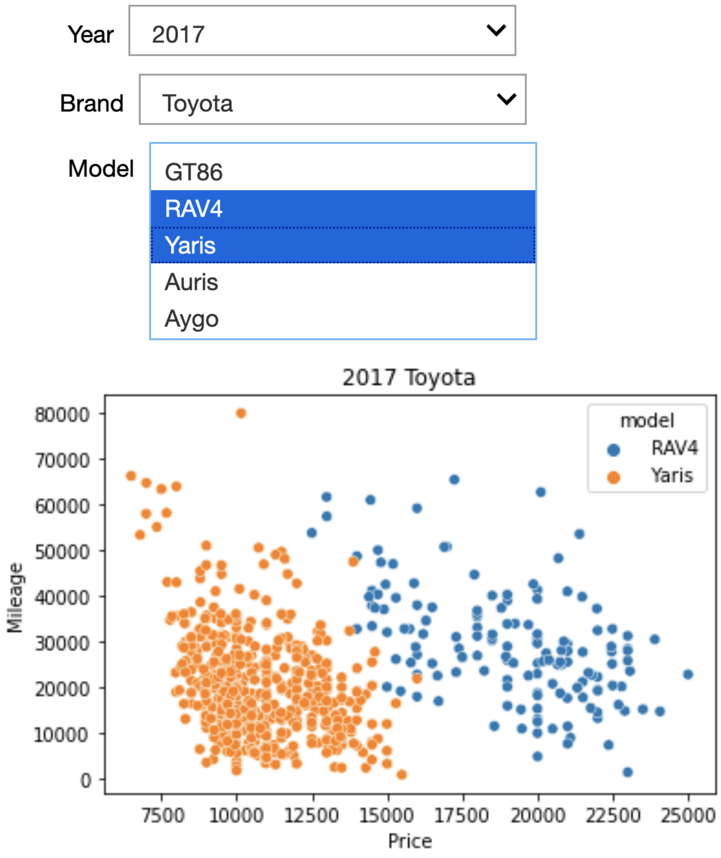
Re-create this interactive visualization using ipywidgets.

The x and y-axes limits are fixed to (20, 140) and (130, 220), respectively, for all sport categories. Make sure the axes are labelled. The colours for the male and female data points should remain consistent as different options are selected. A reminder that this is for gold-medal winning athletes only in the sport categories identified above, for the five Olympics from 2000 to 2016.

6934 Students Only – Question 3: (30 pts)

The image below shows an interactive visualization of used car sales data from the `used_cars.csv` data set. The price of the car is along the x-axis and the mileage of the car along the y-axis.

There are three interactive controls. The first is a dropdown menu in which the user can select one of the available years in the data set. The second is a dropdown menu from which the user can select a car brand based on the brands that were sold in the selected year. The third is a multi-select box that allows the user to select multiple models of cars that were sold for the selected year and car brand.



Year 2017

Brand Toyota

Model GT86 [✓ RAV4] [✓ Yaris] Auris Aygo

Re-create the above interactive experience using ipywidgets.

The brand dropdown menu should be aware of which year was selected in the year dropdown menu, such that the available options for car brand are restricted to only the car brands sold within that year.

Similarly, the car model multi-select box should also be aware of the brand and year selection, so that the available models to select are restricted to the models sold for the selected brand within the selected year.

Submission

Submit your Jupyter notebook (`.ipynb`) through Brightspace.

Late submissions will be subject to a 10% penalty for each hour past the deadline.

Attribution

Submissions should include an attribution section indicating any sources of material, ideas or contribution of others to the submission.

Submissions must represent your independent work.

You are encouraged to use any resources to help with your solution, but your solution must represent independent work. If your submitted work includes unacknowledged collaboration, code materials, ideas or other elements that are not your original work, it may be considered plagiarism or some other form of cheating under MUN general regulations 6.12.4.2 (4.12.4.2 for graduate students) and academic penalties will be applied accordingly.

Avoid academic penalties by properly attributing any contribution to your submission by others, including internet sources and classmates. This will also help distinguish what elements of the submission are original. You may not receive full credit if your original elements are insufficient, but you can avoid penalties for plagiarism or copying if you acknowledge your sources.

Github

I encourage you to store and version your work on GitHub. It is good practice to do so as everyone uses git in the real world.

However, **it is a requirement that git repositories containing assignment material be private.** University regulations (undergraduate 6.12.4.2 and graduate 4.12.4.2) consider it cheating if you allow your work to be copied. There will be zero tolerance for this.