Data Visualization Concepts

BINF4234

Prof. Dr. Renato Pajarola

Exercise and Homework Completion Requirements

1. Exercises and reading assignments are **mandatory** and they must be completed successfully to finish the class and get a sufficient passing final grade.
2. Exercises are graded coarsely into categories **pass** or **fail**.
   - *A **fail** is given to failed submissions and incomplete solutions, and no points are awarded.*
   - *A **pass** indicates that the exercise is sufficiently good to receive the corresponding points.*
   - *Late submissions (up to one day) will result in "-1" point.*
3. The five exercises give rise to the following point distribution: 2 – 3 – 5 – 5.
   - *A **minimum of 7 points** from all four exercises must be achieved to pass the module. Failure to achieve this minimum will result in a failing grade for the entire module.*
   - *Thus at least two exercises have to be correctly solved, and one has to be from the more advanced ones.*
4. We give **bonus points** for students who have completed more than 8 points from all the exercises.
   - *Thus **7 points** from the exercises is required, **8 points** is still normal passing, and **9 and above** would give 1 or more extra bonus points.*
   - *Only the bonus points can and will be added directly to the final grade.*
5. Do not copy assignments, tools to detect copying and plagiarism will be used.
   - *The exercise results are an integral part of the final course grade and therefore the handed in attempts and solutions to the exercises **must be your personal work**.*

Submission Rules

- Submitted code must compile and run without errors using the indicated Python environment, using the included libraries, packages and frameworks. If additional libraries/packages are needed, please specify in a 'readme.txt' file together with your submission.
- The whole project source code must be zipped and submitted before the given deadline, including the output results (saved in .html file or as a screenshot picture).
- Submit your .zip archive named *dvc_ex1_MATRIKELNUMBER.zip* (e.g. dvc_ex1_01234567.zip) through the OLAT course page.
- **Deadline is Sunday, 8 May 2022 at 23:59h**

Exercise 3

The aim of this exercise is to create linked plots using Bokeh visualization techniques. To be specific, you are expected to create two plots:

1. A scatterplot, that shows all the trips in 6 months and their corresponding trip duration. You must also show the vendor id of each trip. This information will be encoded to color visual attribute. Size of the circles must encode the number of passengers for that trip.
2. A histogram of the trip duration information. In other words, trip duration attribute must be divided into bins, then you must show how many trips fall in to each bin.

Task1: Data Preprocessing and scatterplot.

T1.1: Remove the outliers based on the trip duration value.

T1.2: Prepare the ColumnDataSource for plotting: follow the task descriptions in the code skeleton and generate a ColumnDataSource which contains all the information you need for plotting.

(Make sure that you have an idea of how the data source should be before plotting. You may want to read the reference link in the skeleton first before starting.)

T1.3: Add hovering tooltips to the plot in order to provide more detailed information.

T1.4: Create the scatterplot.

Task2: Plot the histogram.

T2.1: Extract the trip duration information from the dataframe and create the histogram for the whole data.

T2.2: Plot the histogram by drawing quads, see the reference link for details.

T2.3: Create two more histogram quads for the selected data, which shall interact with the lasso and box selection widgets.

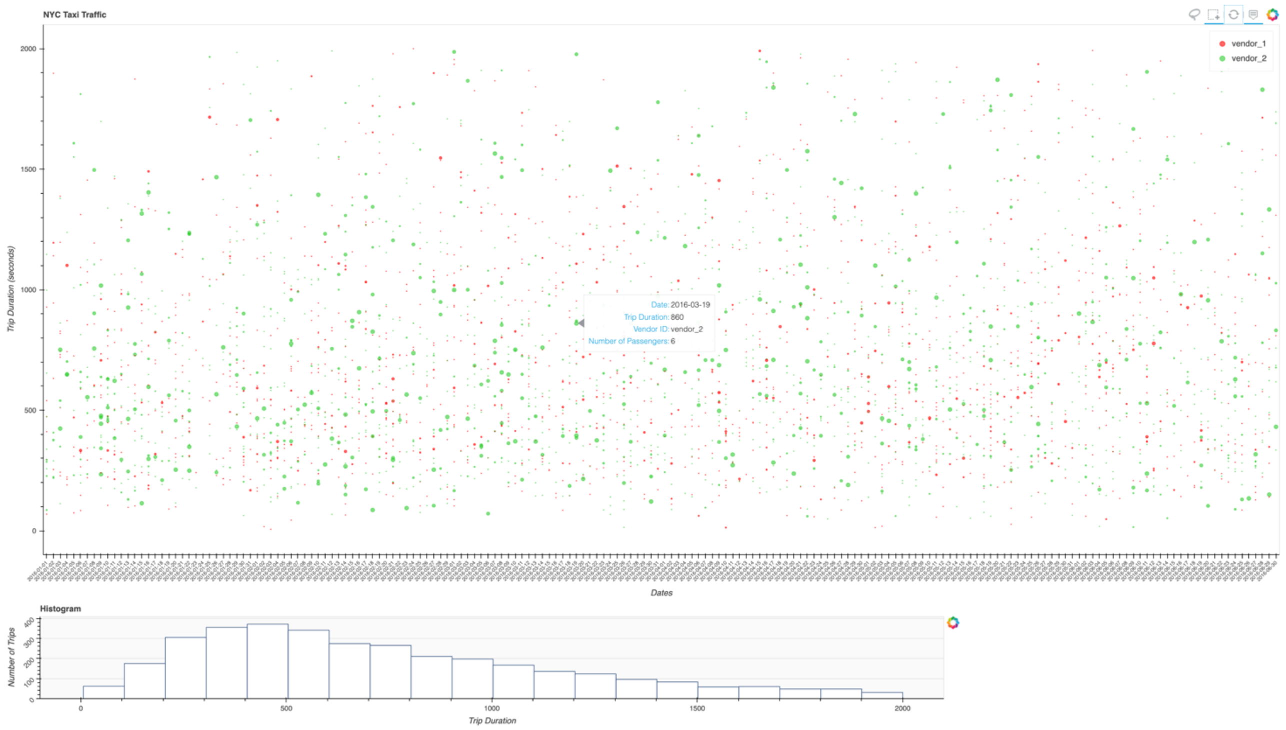
Task3: Create the update function for the Selection Widgets.

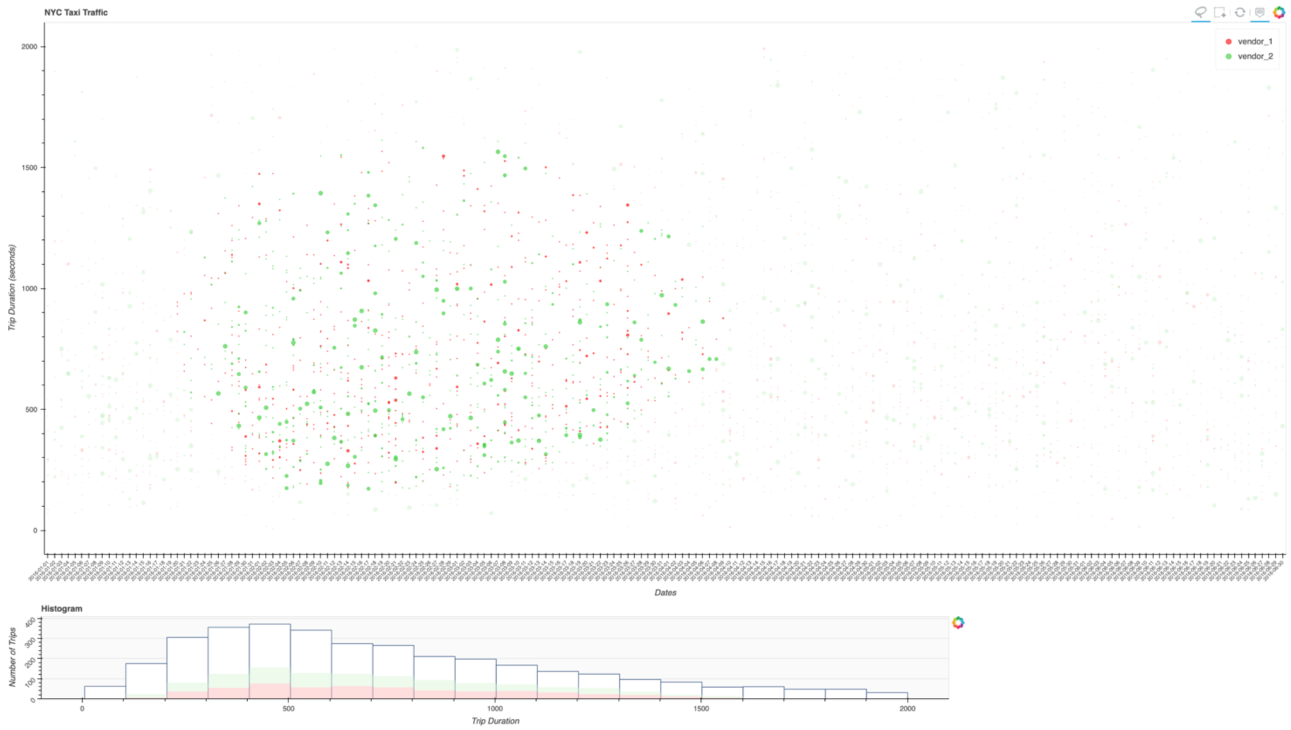
T3.1: Implement the update function that will be triggered when the **lasso** or **box selection** tool is used.

T3.2: Inside of the update function, you need to first extract the data of different vendors from the selected data. Then you can compute the histogram for each vendor, and finally update the two histograms you created in T2.3 with the new hist info. **Attention**, the bottom and top of quads for these two different vendors are different.

T3.3: Save the plot as a .html file.

The following pictures are an example for the desired (but not necessarily the same) visualization result:





Remarks:

- In general, the code skeleton is well structured and divided into groups based on the tasks. However, you may want to change the structure of the skeleton for readability reasons of your own code.

- We recommend to use Jupyter Notebook for your implementation as it can visualize the intermediate output which helps for debugging. However, the final delivery of your code should be .py file rather than .ipynb.

- Try to make good use of the hints and references provided in the skeleton code. (**very important**)

- Try to google first for any Python related issues/bugs.

- Due to the special situation, we don't arrange in person meeting in this semester. Please contact the TA **Emine Didem Durukan (eminedidem.durukan@uzh.ch)** for technical questions regarding the exercise only if needed.

- More than one day late submission will not be accepted and graded.

- The deliverables of this exercise will be a clean version of your code with proper comments, any additional files necessary for executing it (for example, the data file), a "readme.txt" file for your comments or remarks (if necessary), dataset, as well as an export of the final output result in .html or .jpg/.png format. The absence of any required deliverable files will automatically lead to a **FAIL**.