Stats 242 Project

EDA and visualization of medium data sets

Zhewen Hu and Clark Fitzgerald

We visualize a sample of the NYC taxi data set.

Medium data EDA visualization shootout

Contenders:

1. R - Good old vanilla R
2. ggplot2 - popular R library
3. htmlwidgets - A line or two of R code is all it takes to produce a D3 graphic or Leaflet map
4. Matplotlib - established Python plotting library
5. Bokeh - newer Python library

Evaluation criteria:

- **Speed** The primary focus. How long to make a single plot?
- **Aesthetics** Do the defaults look reasonable or are additional tweaks needed?
- **Code Readability** How expressive is the code? Extensibility? maintainability? Learning curve? -___Presentation of the outputs___It is easy to show the output in different platforms?

Timings - All plots should be created on one machine (one of the dept servers) at a time when it's not loaded. Need to make multiple timings.

Save timing data to a CSV file:

program	task	time(seconds)
R	histogram	3.61
...		

Task 1: Histogram

Plot a histogram of the single variable `total_amount` for values of `total_amount` less than 100. Save the result to `histogram.png`.

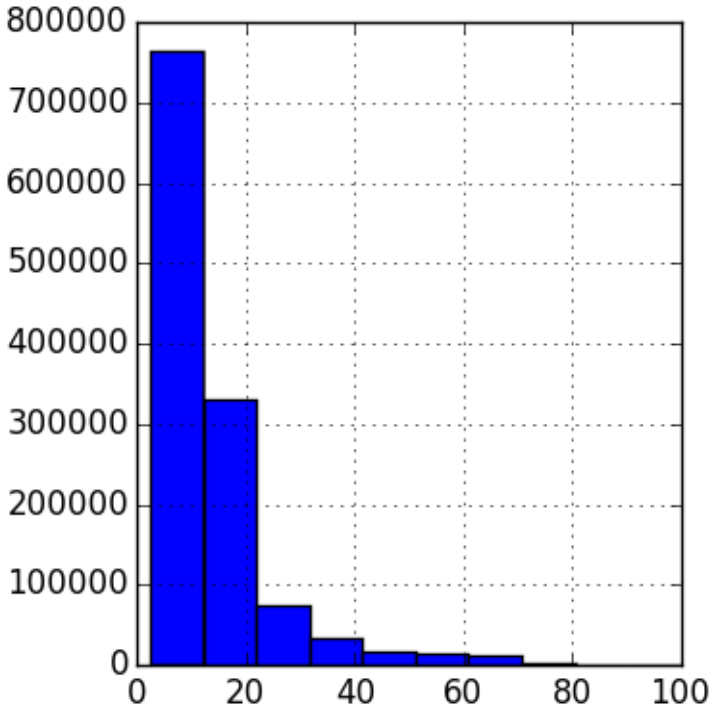

Figure 1: The defaults are not so wonderful

Task 2: Alpha shading

Scatter plot of two variables: trip time in minutes and `total_amount` where the points are semi transparent. This shows the distribution of many points without completely overplotting.

Convert `trip_time_in_seconds` to minutes by dividing by 60.

1. Save the result to `alpha.png`.
2. Filter for rides less than 1 hour and total amount less than 100. Save the result to `alpha2.png`.

Task 3: Sampling

Perform the same scatter plot as the alpha shading, but instead of plotting all points choose a random sample without replacement of 300 points. Save result to `sample.png`.

Task 4: Boxplots

Boxplots of `total_amount` grouped by `payment_type` where `total_amount` is less than 100. Save result to `boxplot.png`.

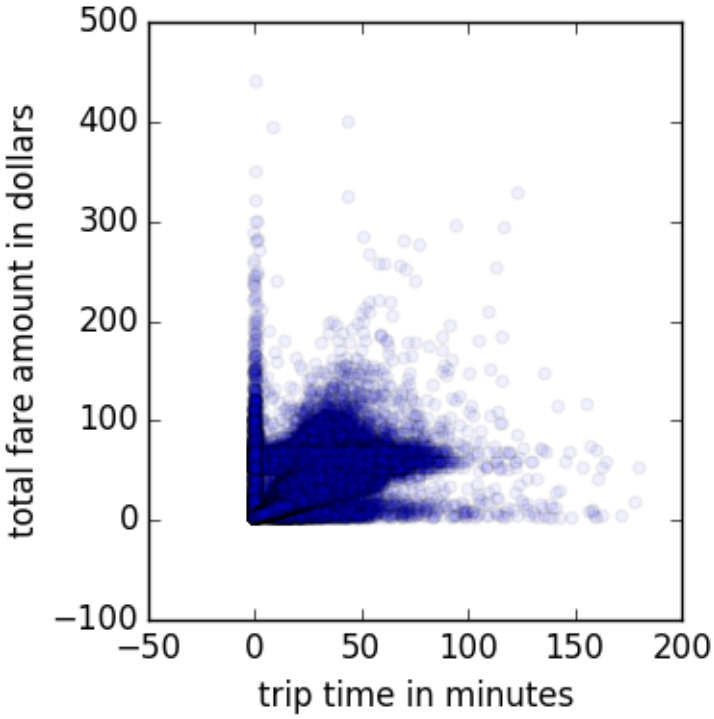

Figure 2: Matplotlib

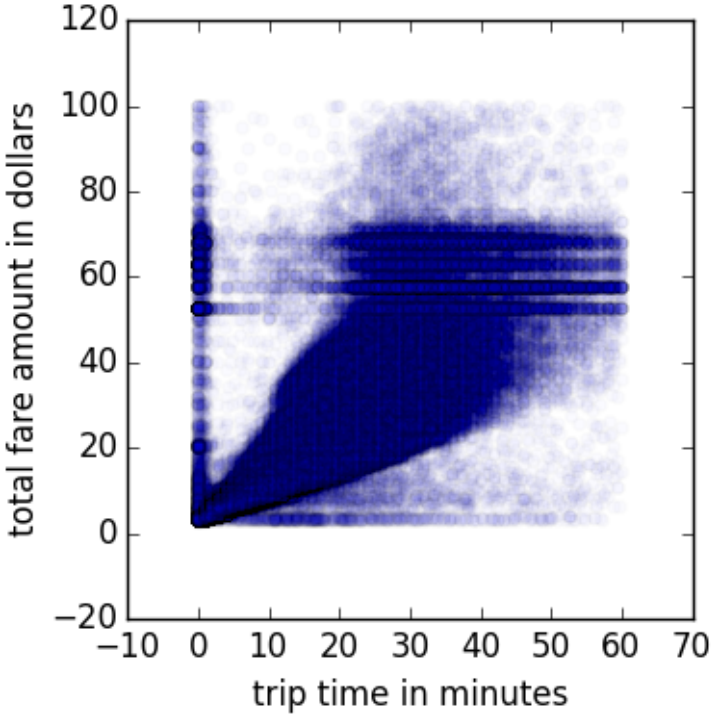

Figure 3: Matplotlib

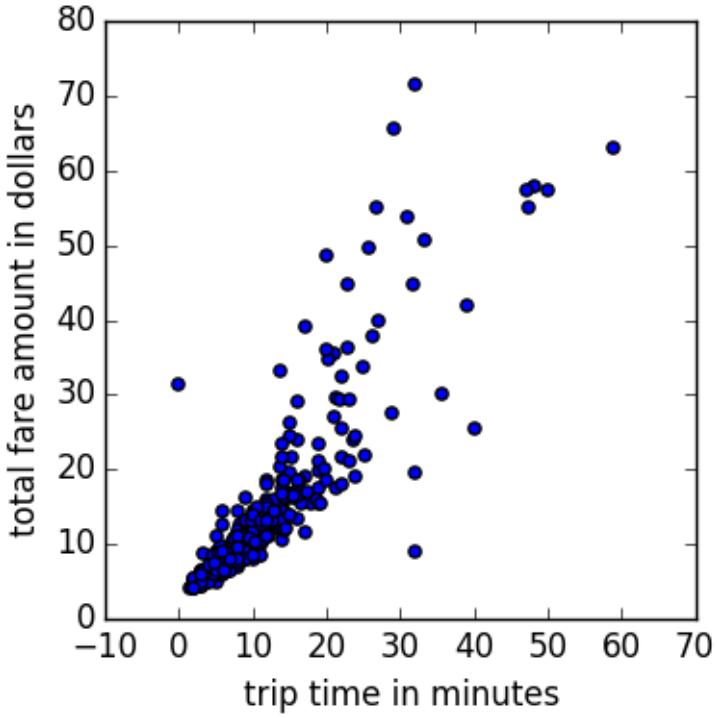

Figure 4: Matplotlib

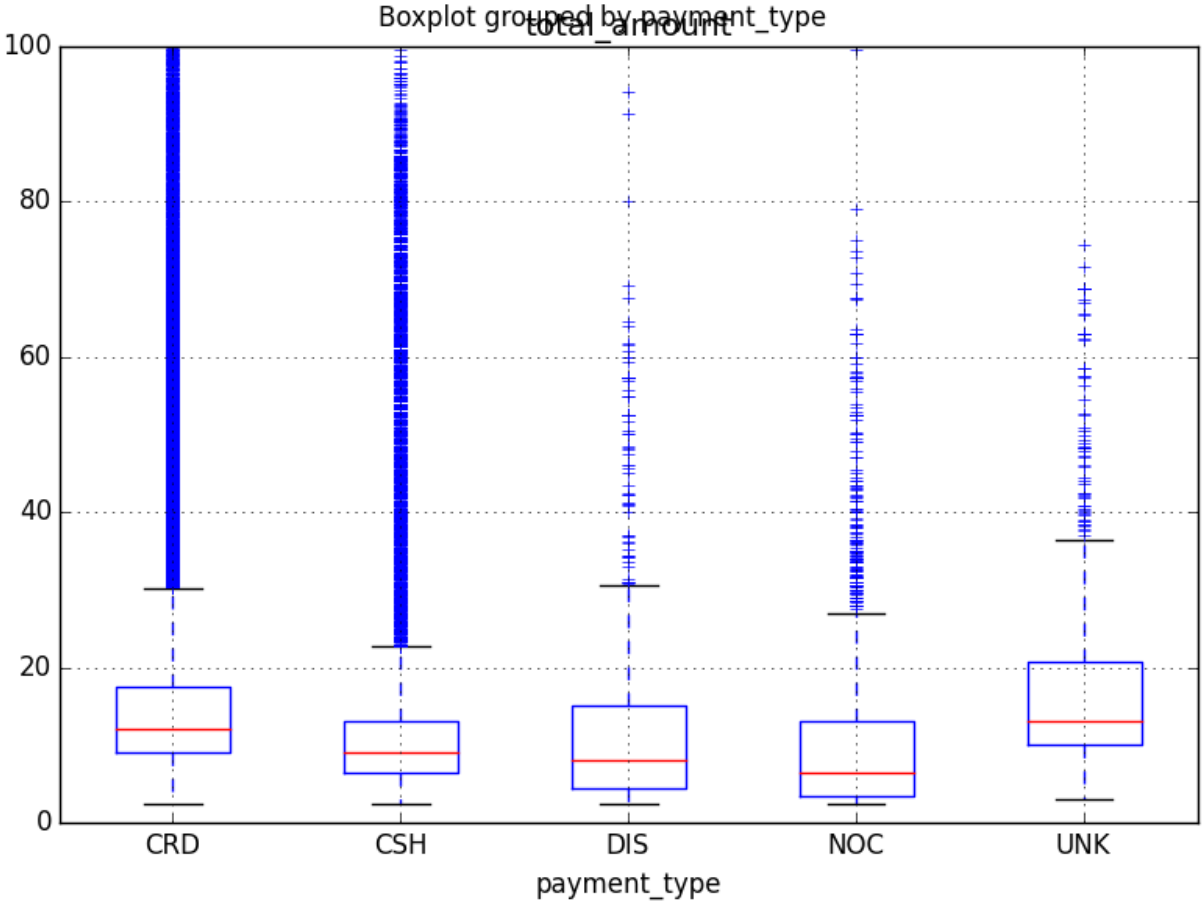

Figure 5: Matplotlib