Week 4 Assignment: Using Plotly for Data Visualization

Overview. In this assignment, you will practice discovering and communicating insights about a set of data by generating a report containing graphical visualizations of that data. This report will challenge you to apply many of the skills you've acquired thus far towards independently interpreting data, including data structuring, exploration, analysis, and presentation.

Goal. By completing this assignment, you will practice and master the following skills:

- Producing data visualizations using the Plotly library or others
- Wrangling multiple data sets of different shapes for specific analyses
- Exploring and drawing conclusions from data
- Learning and utilizing external libraries

Dataset—Airline Delay Causes. You are provided with a data file consisting of on-time-statistics for airlines in the United States. This data set is freely available from The U.S. Department of Transportations: http://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp. A copy of the data containing entries ranging from Feb 2009 – Feb 2019 is provided: data/assignment-airlinedelay.csv

Delays can be categorized into:

- **Air Carrier.** The cause was within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- **Extreme Weather** such as tornado, blizzard or hurricane.
- **National Aviation System (NAS)** which refers to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- **Late-arriving aircraft** where a previous flight with same aircraft arrived late, causing the present flight to depart late.
- **Security** caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

For more details, see RITA -Understanding flight delays and cancellations: (https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations)

Each entry in the airline delay database represents the aggregate monthly data for a particular airline at a particular airport. For example, one entry might be the delay information for Southwest Airlines at St. Louis Airport during December 2013.

Also, each entry lists the number of flights, the number of flights delayed (15 minutes beyond the scheduled arrival time), the number of flights canceled and diverted, the minutes of delay due to carrier delay, weather delay, national air system delays, security delay. The _ct fields list the count of flights experiencing each delay (they add up the arr_del15 column). Because a flight can have multiple delay types, these numbers are real numbers. For example, if a flight was 30 minutes late, 15 minutes due to weather and 15 minutes due to security it will contribute .5 to weather_ct and .5 to security_ct. It will also contribute 15 to the weather_delay field and 15 to the security_delay field.

Finally, the arr_delay column is the sum of total delay minutes for the document (i.e., arr_delay = carrier_delay+ weather_delay + nas_delay + security_delay + late_aircraft_delay)

The names of each field in an entry:

Label/field name	Meaning
year	The year the month occurred
month	The month for which data was collected
carrier	The airline code
carrier_name	The airline name
airport	The airport code
airport_name	The airport name
arr_flights	# of flights that arrived at the airport
arr_del15	# of flights that arrived >= 15 minutes late
carrier_ct	# of flights delayed due to the carrier
weather_ct	# of flights delayed due to weather
nas_ct	# of flights delayed due to national air system
security_ct	# of flights delayed due to security
late_aircraft_ct	# flights delayed because a previous flight using the same aircraft was late
arr_cancelled	# of canceled arrivals
arr_diverted	# of scheduled arrivals that were diverted
arr_delay	Sum of the delay minutes
carrier_delay	Total minutes of delays due to carriers
weather_delay	Total minutes of delays due to weather
nas_delay	Total minutes of delays due to natl. air service
security_delay	Total minutes of delays due to security
late_aircraft_delay	Similar to late_aircraft_ct. The total minutes of delay due to a previous flight using the same aircraft arriving late.

This is an open-ended assignment requiring your critical thinking skills. Be ready to explore, experiment, and program on your own! You can visualize the flight delay data with maps, bar charts, line charts, scatterplots, and many others.

Some guidance questions include:

- Which airline should you fly on to avoid significant delays?
- Which months should you fly on to avoid significant delays?
- Any relationships between seasons and flight delays?
- Which region of airport has the most significant number of delays?
- What is the major reasons for delay?

Submission.

You should work individually on this assignment.

Turn in one zip file with the following components:

- Images of the visualizations (.png or .jpg)
- Your code (with any text editor in the format of .py or .ipynb so others can run your code and see the interactivity).

The Rubric. We thought it might be helpful to have some sort of guidance when reviewing everyone's data visualization. The rubric is based on effectiveness on three areas: coding, presentation, and storytelling.

When grading, you do not need to respond to every point in the rubric, but instead use the rubric as a guideline for the peer review.

Rubric

We’d like you to assign numerical scores (from 1-5) for each of these questions. Full score is 50.

	Needs work: 1	**2**	**Satisfactory: 3**	**4**	**Excellent: 5**
Coding					
Coding style	Many errors in coding style, little attention paid to making the code human readable		Coding style lacks refinement and has some errors, but code is readable and has some comments		Student has gone beyond what was expected and required, coding manual is followed, code is well commented
Coding strategy	Code tackles complicated problem in one big chunk. Code is repetitive and could easily be functionalized. No anticipation of errors.		Code is correct, but could be edited down to leaner code. Some "hacking" instead of using suitable data structure. Some checks for errors.		Complicated problem broken down into sub-problems that are individually much simpler. Code is efficient, correct, and minimal. Code uses appropriate data structure (list, data frame, vector/matrix/array). Code checks for common errors
Presentation					
Graph type	Graph(s) poorly chosen to support questions.		Graph(s) well chosen, but with a few minor problems: inappropriate aspect ratios, poor labels.		Graph(s) carefully tuned for desired purpose. One graph illustrates one point
Grouping and Search	Difficult to find groups, follow lines, etc. Elements to be compared are not aligned. Distinct variables are mapped to integral dimensions (e.g. point width and height). Distinct elements cannot be discriminated.		The visualizations show attempts to convey information about the data through gestalt and preattentive processing features. But could be made more so.		Gestalt and preattentive processing features are chosen to ease task (find important groups, follow lines, etc.) Elements to be compared are aligned, as much as possible. Distinct variables are mapped to separable dimensions. Choice of colors, shapes, etc. is easy to discriminate.

Annotation and textual summaries	No message, or message is not at all supported by text and graphs The visualizations are Difficult to navigate and interpret, cluttered/ complicated.		Simplistic but clear message. Text and graphs support the message adequately		The visualizations include titles, appropriate labeling, textual summaries. Navigation is guided. Clear, multifaceted message. Titles, labeling, textual summaries,and annotations are used to support the message strongly.
Color and Font	Poor Contrast (colors or fonts are not clearly distinguished, or are not used to provide structure)and poor Repetition (too many different colors or fonts, and do not convey unity).		Effective use of either Contrast or Repetition, though not both.		Shows strong use of principles of Contrast and Repetition: limited color palette and font choices convey structure and unity.
Layout	Poor Alignment (elements are placed haphazardly)and poor Proximity (no obvious grouping ofelements or separation of groups; unclear whichcaptions match which text/graphs).		Effective use of either Alignment or Proximity, though not both. Alignment is adequate within sections of graphic, but not across them.		Shows strong use of principles of Alignment and Proximity: bold alignment guides reader through the graphic; proximity groups related elements together and separates distinct groups.
Storytelling					
Informative	Graph highlights no interesting or useful pattern. Pattern is not indicated in title or not described in write		Graph highlights some interesting or useful pattern but readers have to mentally compute differences instead of seeing them directly.		Graph clearly highlights any trend or pattern in the data, which is indicated in the title and described in the writeup.

	up. Readers have to mentally compute differences instead of seeing them directly.				Interesting differences or comparisons are plotted directly
Comprehensible	Graphic has no (or unclear) title, axis labels, or legend. Axis ticks are unmarked or are marked at arbitrary, unhelpful numbers.		Graphic has a clear title, axis labels, or legend, but axis ticks are unmarked or are marked at arbitrary, unhelpful numbers.		Graphic has an informative title, axis labels, and legend(if relevant). Axis ticks are labeled with sensible, round numbers.
Other	Poor craftsmanship with many imperfections. Does not cite data sources.		Shows good craftsmanship and use of computer layout. Cites all data sources.		Shows decent craftsmanship on computer layout, with minor imperfections. Cites data sources.