Line and Scatter Plots

Lab Aims

In this lab, you will use your new knowledge to propose solutions to real-world scenarios. To succeed, you will need to import data into Python, answer questions using the data, and data visualization tasks such as generate line charts to understand patterns in the data.

1 What have you learned in this Lab?

In this lab, you will learn how to create many different advanced chart types. Now, some quick commands that you can use to do these charts.

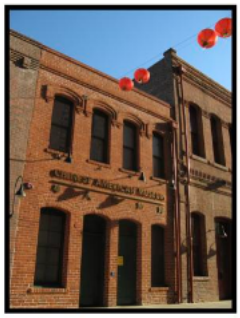
- *sns.lineplot* - Line charts are best to show trends over a period of time, and multiple lines can be used to show trends in more than one group.
- sns.scatterplot - Scatter plots show the relationship between two continuous variables; if color-coded, we can also show the relationship with a third categorical variable.
- *sns.regplot* - Including a regression line in the scatter plot makes it easier to see any linear relationship between two variables.
- *sns.histplot* - Histograms show the distribution of a single numerical variable.
- *sns.countplot* Show the counts of observations in each categorical bin using bars.
- *sns.barplot* - Bar charts are useful for comparing quantities corresponding to different groups.

2 First Scenario

You have recently been hired to manage the museums in the City of Los Angeles. Your first project focuses on the four museums pictured in the images below.

Avila Adobe

Firehouse Museum



Chinese American Museum

America Tropical Inventive Center

You will leverage data from the Los Angeles Data Portal that tracks monthly the number of visitors to each museum.

2.1 Load the data

- Your first assignment is to import and configure the Python libraries that you need to complete this exercise.
- Then, read the LA Museum Visitors data file ('museum_visitors.csv') into `museum_data` data-frame. In addition, the name of the column to use as index is "Date".

2.2 Review the data

In this part, we will review the dataset, in order to first understand what you have "in your hand"

- What is the shape size of the `museum_data` dataset?
- Print the first and last 7 rows of the `museum_data` data frame.
- List the names of the features included in this dataset and their data type.
- Check for and remove all rows with NaN, or missing, values in the `museum_data` Data-Frame.
- Check for and remove all duplicate rows in the student Data-Frame.

3 Summarize the Data

- Provide a summary statistic, which includes the mean value, the minimum value and the maximum value for each feature, as well as its standard deviation, and the 25%, 50% (i.e., the median) and 75% percentiles.

Convince the museum board

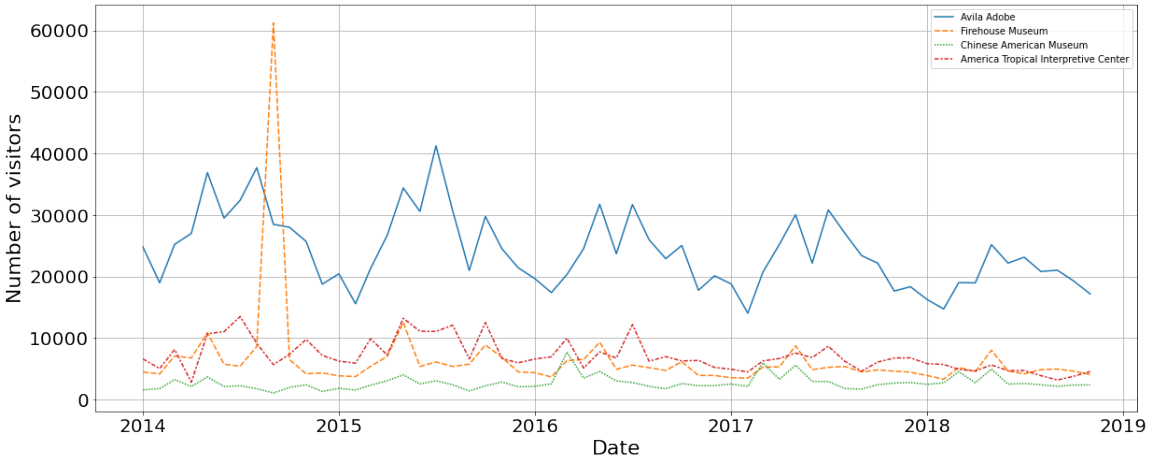
The Firehouse Museum claims they ran an event in 2014 that brought an incredible number of visitors, and that they should get extra budget to run a similar event again. The other museums think these types of events aren't that important, and budgets should be split purely based on recent visitors on an average day.

Therefore, to show the museum board how the event compared to regular traffic at each museum, create a line chart that shows how the number of visitors to each museum evolved over time. Your figure should have four lines (one for each museum) as indicated in the following.

3.1 Assess seasonality

When meeting with the employees at **Avila Adobe**, you hear that one major pain point is that the number of museum visitors varies greatly with the seasons, with low seasons (when the employees are perfectly staffed and happy) and also high seasons (when the employees are understaffed and stressed). You realize that if you can predict these high and low seasons, you can plan ahead to hire some additional seasonal employees to help out with the extra work.

- Create a line chart that shows how the number of visitors to Avila Adobe has evolved over time.
- Does Avila Adobe get more visitors:



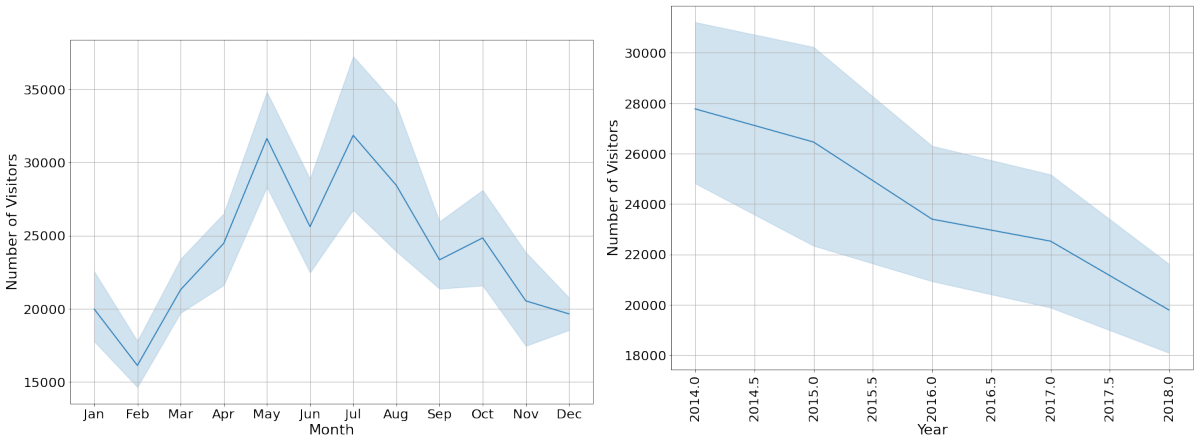
  - in September-February (in LA, the fall and winter months), or
  - in March-August (in LA, the spring and summer)?

  You might need to use the following code to answer this question:

```
import matplotlib.dates as mdates
years = mdates.YearLocator()   # every year
months = mdates.MonthLocator()  # every month
years\_fmt = mdates.DateFormatter('%Y-%m')
fig, axes = plt.subplots(figsize=(20,7))
.
axes.xaxis.set\_major\_locator(months)
axes.xaxis.set\_major\_formatter(years_fmt)
axes.xaxis.set\_minor\_locator(months)
```

- Using the extracted information, when should the museum staff additional seasonal employees?
- Plot the variation in visitor numbers of the **Avila Adobe** museum by month name, and another plot by year. The obtained results are presented in Figure **??**:

What you can conclude from these results? Are these findings similar to those of other museums?

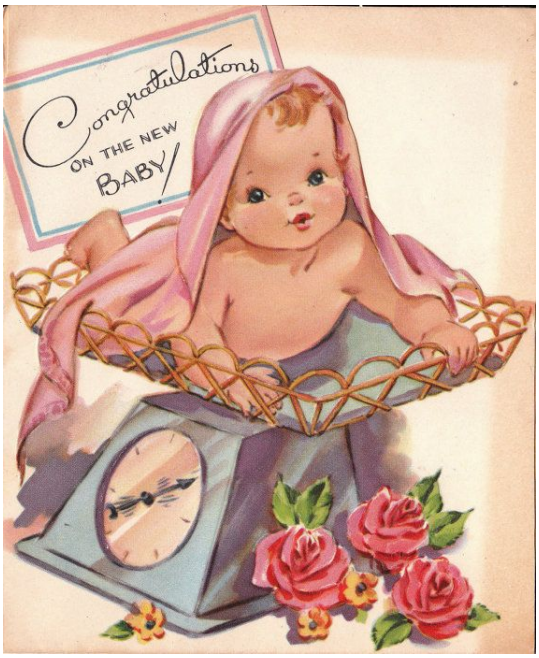


((a)) Per month name ((b)) Per year

Figure 1: Variation of the number of visitors per month (a) and per year(b)

Second Scenario

In fact, we will practice data visualization using data on births from the state of North Carolina. The data set ("ncbirth.csv") that shows up in your Environment is a large data frame. Each observation or case is a birth of a single child. In addition, this data set contains 800 observations (rows or cases) and 13 variables (columns).



The variable:

- "weeks" represents the pregnancy duration.
- "weight" represents the weight of the baby at birth in pound

- ”premie” indicates whether a birth was early (premie) or went full term.
- ”mature”: the maturity status of a mother
- ”gained”: weight gained by mother during pregnancy in pounds.
- ”gender”: gender of the baby, female or male.
- ”habit”: status of the mother as a nonsmoker or a smoker.

Tasks

1. Writing the necessary code to load the data.
2. Analyze the relation between ”weeks” and ”weights” using the required function of the ”seaborn” library. Include axis labels with measurement units, and a title. Is there a positive or negative relationship between these variables?
3. Make a graph showing weeks again on the x axis and the variable ”gained” on the y axis (the amount of weight a mother gained during pregnancy). Include axis labels with measurement units, and a title.
4. Visually assess the association between the babies’s weight (weight) and the mother weight gain (gained) ?
5. Color the points of the previous plot based on the the ”premie” variable. How many variables are now displayed on this plot? what is your conclusion?
6. Make a new scatter plot that shows a mothers age on the x axis (variable called mage) and birth weight of newborns on the y axis (weight). Color the points on the plot based on the gender of the resulting baby (variable called gender). Does there appear to be any strong relationship between a mother’s age and the weight of her newborn?
7. Plot the distribution of the ’habit’ feature (mother is a nonsmoker or a smoker) split by the maturity status of the mothers (mature). What you can conclude?
8. Plot the distribution of pregnancy duration (variable called weeks) split by the premature classification (premie or full-term). Add a title and axis labels. The y axis is labeled count. What you can conclude?
9. Make a distribution of newborn birth by gender of the child split by the maturity status of the mothers (mature). What you can conclude?