Final Exam- Fall 2019

DSO 545: Statistical Computing and Data Visualization

12/17/2019

Instructions

- This is an open notes exam (2 double-sided paper). You are NOT allowed to use the Internet as a resource except for webscrapping, downloading the files from blackboard, and uploading your answer files back to blackboard.
- Answer all questions below
- Don't change the data file names
- You have 120 mins to finish this exam
- You are NOT allowed to communicate with ANY PERSON in or outside the class during the exam period
- Suggested **max** times to spend on each case:

Download Data	Case 01	Case 02	Case 03	Case 04	Upload Files
5 mins	20 mins	20 mins	20 mins	50 mins	5 mins

"I hereby certify that I have adhered to the university policies regarding ethical behavior in preparing for and completing this midterm exam. I will not discuss the exam questions and solutions with anyone in the classroom or outside the classroom via any means before Dec 18, 2019."

Name : ----------

Signature : ----------

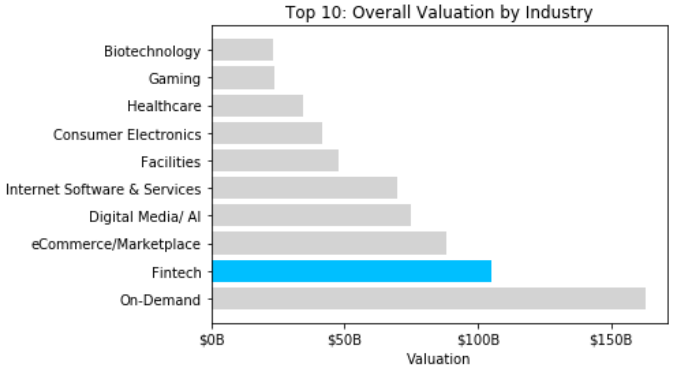
Case 01: Data Manipulation & Visualization

A unicorn startup or unicorn company is a private company with a valuation over $1 billion. As of January 2019, there are more than 300 unicorns around the world. In this **exercise**, we will study these unicorns, their valuations, industry, as well we the top investors.

Datasets: *unicorn.csv*

- **Company**: The name of the unicorn company
- **Valuation**: The valuation (in billions $) of the unicorn company
- **DateJoined**: Date in which the company became a unicorn
- **Country**: Location of the Company
- **Industry**: To which industry does the company belong?
- **SelectedInvestors**: A list of top investors in the unicorn

(1) *(2 points)* Create a barchart that shows the overall valuation of the unicorn industries. (use *plt.barh()* to create a horizontal barchart. It takes the same arguments as *plt.bar()*). The color for the 'Fintech' bar is '*deepskyblue*', and the rest of the bars is '*lightgray*'.

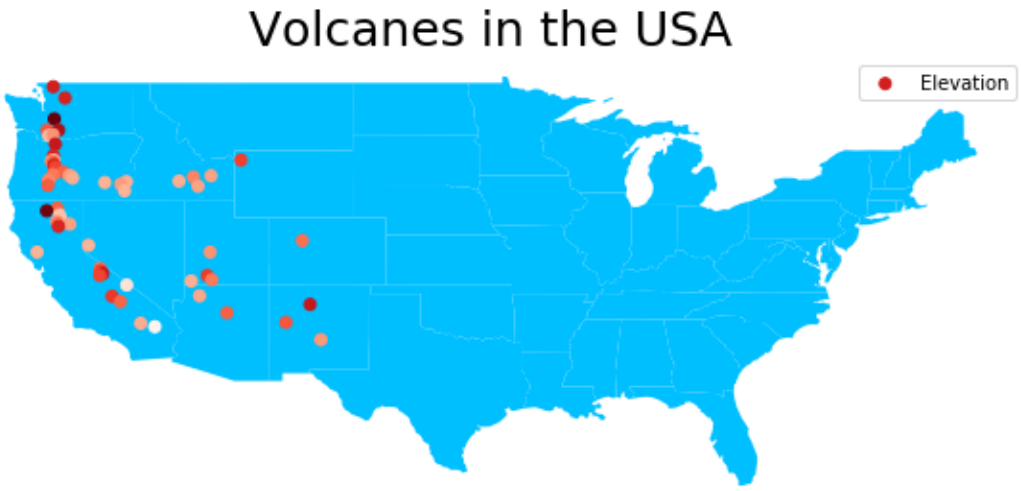


Case 02: GeoMaps

In this case, we will use a dataset that shows the locations of all volcanoes in the USA.

Datasets: *volcanoes_usa.csv* & *us-states.json*

(2) *(2 points)* Create a Choropleth map that shows the location of the volcanoes in the USA. The color intensity of the dots on the map represent the elevation of these volcanoes. (Use cmap = 'Reds' for the color shades of elevation, and color = 'deepskyblue' for the base map of the USA)



Case 03: Regular Expressions

Have you seen the TV show *Friends*? In season 8 episode 22 of Friends, one of the characters- Joey, is figuring out how much money he owes another character in the show-Chandler, for rent, acting lessons, dance lessons, head shots, etc.

(3) *(2 points)* How much did Joey owe Chandler?

Datasets: `*friends.txt*` summarizes a conversation between Joey and Chandler.

Hint:

- The amounts have dollar signs preceding them
- The $ sign is a wild character, so to refer to it in a regular expression to represent a dollar sign, you will need to escape it using the escape character "\" as follows \$
- To read a text file in Python, you can use the following code:

```
# Open a file: read only
file = open("friends.txt","r")

# read all lines at once
friends = file.read()

# close the file
file.close()
```

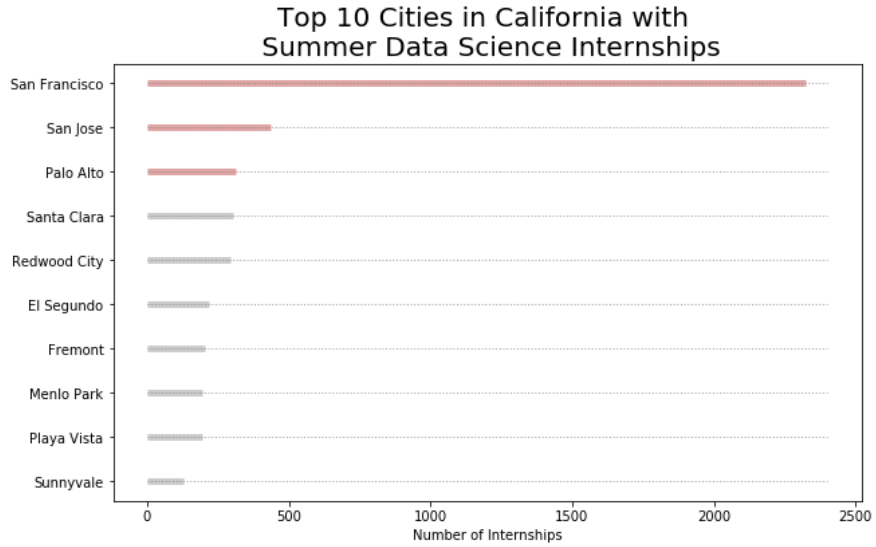
Case 04: Web Scrapping

Each one of you will be looking for an analytics/data science internship for Summer 2020. In order to automate your search process, your job is to build a web scrapper that searches for all internships related to data science, and extracts for each position the company name, company location, job title, and job description link.

Before you scrape the data yourself, you will first work with the dataset provided in the file ("datascience_internships.csv") and then in the second part you will scrape it using Python.

Datasets: *datascience_internships.csv*

(4) *(2 points)* Read the datasciecne_internships.csv file and show the top 10 cities in California with the most internships in data science as follows. The color of the top 3 cities with most internships is "firebrick" and the rest are "grey". The width of the line bars is 5, and the alpha level of these line bars is 0.7.



(5) *(2 points)* Use the following link to scrape and extract for each position the **company name**, **company location**, **job title**, and **job description link**.

Link: https://www.indeed.com/jobs?q=data+science+intern&l=California