Final

BDA 503 - Fall 2018

General Instructions

Your take home final consists of 3 parts. First part is about some simple questions and their answers. These questions might include coding, brief comments or direct answers. Second part is about your group projects. You are asked to make a contribution to your project report with an additional analysis with two/three visualizations. Third part is about gathering real life data and conducting analysis on it.

Here are significant points that you should read carefully.

- Your final starts on December 28, 2018; 11:00, and ends on January 7, 2019; 23:59. Late submissions are not accepted, strict (meaning you will get 0 points). Estimated workload is 1 day, 2 days tops. If it gets longer, blame is on you (you probably overthink about it).
- Your main submissions will be through Blackboard, no email. Please refrain from posting on your progress journals until January 8, 2019. After then, it is appreciated.
- You will submit RMarkdown generated pdf files with code (unless stated otherwise). You will submit only a single pdf containing all 3 parts.
- All works should be individual and original. (Single exception: On data gathering in part 3, you can work together and refer to the same RData file.)
- Instructor support will be minimal. I will try to answer technically ambiguous points but I will generally not respond to consulting questions (e.g. "Am I doing it ok?" You probably are, given your overall performance). Questions are designed to measure your opinions and I don't want to color your perspective.

Part I: Short and Simple (20 pts)

The purpose of this part is to gauge your apprehension about data manipulation, visualization and data science workflow in general. Most questions have no single correct answer, some don't have good answers at all. It is possible to write many pages on the questions below but please keep it short. Constrain your answers to one or two paragraphs (7-8 lines tops).

1. What is your opinion about Python vs R debate? To what extent do you agree with the post on https://www.dataschool.io/python-or-r-for-data-science/? Be honest, you won't be penalized or rewarded for stating your opinions; only by the quality your arguments.

2. What is your exploratory data analysis workflow? Suppose you are given a data set and a research question. Where do you start? How do you proceed? For instance, you are given the task to distribute funds from donations to public welfare projects in a wide range of subjects (e.g. education, gender equality, poverty, job creation, healthcare) with the objective of maximum positive impact on the society in general. Assume you have almost all the data you require. How do you measure impact? How do you form performance measures? What makes you think you find an interesting angle?

   Would you present an argument for a policy that you are more inclined to (e.g. suppose you are more inclined to allocate budget to fix gender inequality than affordable healthcare) or would you just present what data says? In other words, would the (honest) title of your presentation be "Gender Inequality - The Most Important Social Problem Backed by Data" or "Pain Points in Our Society and Optimal Budget Allocation"?

3. If you had to plot a single graph using the `flights` data what would it be? Why? Make your argument, actually code the plot and provide the output. (You can find detailed info about the movies data set in its help file. Use `?flights`, after you load `nycflights13` package.)

```
library(dplyr)
library(nycflights13)
glimpse(flights)
```

```
## Observations: 336,776
## Variables: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...
```

Part II: Extending Your Group Project (30 pts)

In this part you are going to extend your group project with a single additional analysis supported by some visualization. You are tasked with finding the best improvement on the top of your group project. About one page is enough, two pages tops.

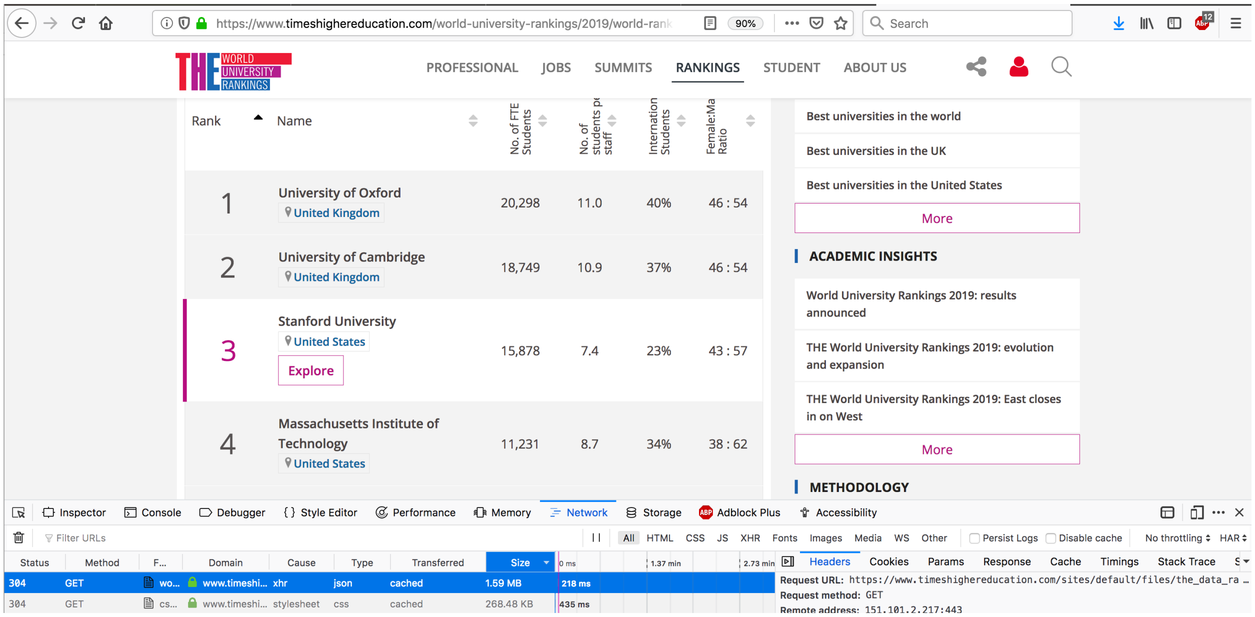
Part III: Welcome to Real Life (50 pts)

As all of you know well enough; real life data is not readily available and it is messy. Also you will face situations where you need to discover and learn another framework. In this part, you are going to gather data from TIMES Higher Education rankings. You can use all the data provided on 2019 Rankings and before. Take some time to see what are offered in the data sets. Choose an interesting theme which can be analyzed with the given data and collect relevant data from the service. Some example themes can be as follows.

- Place of Turkish universities and underlying differences in categories.
- Distribution of gender parity and rankings.
- Correlation of scores.
- Distribution of countries in different rank brackets and under different categories.

a) Gather the data, bind them together and save in an .RData file. You can make .RData file available online for everybody. Provide the data link in your analysis. You can work together with your friends to provide one comprehensive .RData file if it is more convenient to you. (You don't need to report any code in this part.)

Tip You might need some help getting the data from the website. Use `jsonlite::fromJSON` with `flatten=TRUE` to the json file served from the website. You can get the JSON address using a modern browser web tools.



https://www.timeshighereducation.com/world-university-rankings/2019/world-rank

Rank	Name	No. of FTE Students	No. of students pe staff	Internation Students	Female:Ma Ratio
1	University of Oxford United Kingdom	20,298	11.0	40%	46 : 54
2	University of Cambridge United Kingdom	18,749	10.9	37%	46 : 54
3	Stanford University United States Explore	15,878	7.4	23%	43 : 57
4	Massachusetts Institute of Technology United States	11,231	8.7	34%	38 : 62

Best universities in the world

Best universities in the UK

Best universities in the United States

More

ACADEMIC INSIGHTS

World University Rankings 2019: results announced

THE World University Rankings 2019: evolution and expansion

THE World University Rankings 2019: East closes in on West

More

METHODOLOGY

Status	Method	F...	Domain	Cause	Type	Transferred	Size	
304	GET	wo...	www.timeshi...	xhr	json	cached	1.59 MB	218 ms
304	GET	cs...	www.timeshi...	stylesheet	css	cached	268.48 KB	435 ms

Request URL: `https://www.timeshighereducation.com/sites/default/files/the_data_ra …`

Request method: GET

Remote address: 151.101.2.217:443

b) Perform EDA on the data you collected based on the theme you decided on. Keep it short. One page is enough, two pages tops. Original and interesting work is important (data sharing is good, but be careful about idea sharing). If you are interested and want to keep going, write a data blog post about it. I will not grade it but I can share it on social media.