

Final

BDA 503 - Fall 2017

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 January 6, 2018; 11:00. It ends on January 9, 2018; 11:00. Late submissions until January 9, 2018; 23:59 (penalty -25 points).
- Your main submissions will be through Blackboard, no email. Please refrain from posting on your progress journals until January 10, 2018. After then, it is appreciated.
- You will submit RMarkdown generated pdf files. 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 two y-axis graphs? Do you use it at work? Is it a good practice, a necessary evil, or plain horrible? See Hadley Wickham’s point (and other discussions in the topic) here before making your argument (<https://stackoverflow.com/a/3101876/3608936>). See an example of two y-axis graph on https://mef-bda503.github.io/gpj-rjunkies/files/project/index.html#comparing__of__accidents__of__departures
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. What are the differences between time series and non time series data in terms of analysis, modeling and validation? In other words what makes Bitcoin price movements analysis different from diamonds (or carat) data set?
4. If you had to plot a single graph using the data below 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 ?movies, after you load ggplot2movies package.)

```
library(ggplot2movies)
movies

## # A tibble: 58,788 x 24
##           title year length budget rating votes   r1   r2
##           <chr> <int> <int> <int> <dbl> <int> <dbl> <dbl>
## 1           $ 1971   121    NA    6.4   348   4.5   4.5
## 2    $1000 a Touchdown 1939    71    NA    6.0    20   0.0  14.5
## 3    $21 a Day Once a Month 1941     7    NA    8.2     5   0.0   0.0
## 4    $40,000 Climax Show, The 1975    71    NA    3.4    17  24.5   4.5
## 5    $pent 2000    91    NA    4.3    45   4.5   4.5
## 6    $windle 2002    93    NA    5.3   200   4.5   0.0
## 7    '15' 2002    25    NA    6.7    24   4.5   4.5
## 8    '38 1987    97    NA    6.6    18   4.5   4.5
## 9    '49-'17 1917    61    NA    6.0    51   4.5   0.0
## # ... with 58,778 more rows, and 16 more variables: r3 <dbl>, r4 <dbl>,
## #   r5 <dbl>, r6 <dbl>, r7 <dbl>, r8 <dbl>, r9 <dbl>, r10 <dbl>,
## #   mpaa <chr>, Action <int>, Animation <int>, Comedy <int>, Drama <int>,
## #   Documentary <int>, Romance <int>, Short <int>
```

Part II: Extending Your Group Project (30 pts)

In this part you are going to extend your group project with an additional analysis supported by some visualizations. 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. In this part, you are going to gather data from Higher Education Council's (YÖK) data service. You can use all the data provided on <https://istatistik.yok.gov.tr/>. 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.

- Gender disparity in the academic faculty.
 - Change in the number of people in different academic positions in years.
 - Professor/student ratios.
 - Capacities in different departments.
 - Comparative undergraduate / graduate student populations.
 - Number of foreign students/professors and where they come from.
- a) Gather the data, bind them together and save in an .RData file. 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.)

- b) Perform EDA on the data you collected based on the theme you decided on. Keep it short. One to two pages is enough, three pages tops. 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.