ADANA ALPARSLAN TÜRKEŞ
SCIENCE AND TECHNOLOGY UNIVERSITY

EEE356 - Data Analytics
Midterm Exam
Prepared by Dr Kasım Zor

19th Apr 2024, 09:15–12:00

Full Name : ____________________ Student ID: ________________

Grade Table (for Lecturer use only)

Question	Points	Score
1	40	
2	60	
Total:	100	

Instructions for Midterm Exam

Welcome to the midterm exam of EEE356 - Data Analytics and good luck!

Please read the following rules and confirm by signing that you have read and understood the rules before you receive your exam:

- The midterm exam shall be conducted between 09:15 and 12:00. Exam duration is 165 minutes. Students must finalise the exam by delivering it before 12:00. Students are not allowed to leave the exam in the first 30 minutes.

- Student ID cards shall visibly be on the edge of desks till the end of the exam. Students without the student ID cards or Turkish identity cards shall not be participated into the exam.

- This is a closed-book exam which means that students are not allowed to take notes, books, or any other reference material into the exam. Throughout the exam, students shall not possess mobile phones and electronic devices that are capable of storing, receiving or transmitting information or electronic signals, such as computerised watches.

- Students are not allowed to take a glance at the exam questions until told to do so. Students shall not communicate with any other student under any circumstances during the exam period. A student, who cheats, tries to cheat during the exam, or is identified to be cheating after investigating exam documents, is given 0 (zero) for that exam and a disciplinary investigation is opened against the student.

- An incorrect answer to a question is awarded no marks with no consideration of any partial credit. Therefore, no partial credit will be given.

In recognition of and in the spirit of the above rules which constitute Adana Alparslan Türkeş Science and Technology University Honour Code, I certify that I will neither give nor receive unpermitted aid on this examination.

Signature: ____________________

1. Answer the following questions.

   (a) **(6 points)** Define the term *Data Analytics* with your own words.

   (b) **(9 points)** Explain the KDD process by listing the main steps.

   1.
   2.
   3.
   4.
   5.
   6.
   7.
   8.
   9.

   (c) **(9 points)** Compare the three most popular data science languages, namely Julia, Python, and R in terms of the type of translator, the computational speed for scientific calculations, and the data visualisation skills.

Feature	**Julia**	**Python**	**R**
Translator	□ Compiler □ Interpreter	□ Compiler □ Interpreter	□ Compiler □ Interpreter
Speed	□ Slow □ Normal □ Fast	□ Slow □ Normal □ Fast	□ Slow □ Normal □ Fast
Visualisation	□ Fair □ Elegant	□ Fair □ Elegant	□ Fair □ Elegant

   (d) **(6 points)** Fill in the blanks with the appropriate words in the following sentences.
   A __________ in data analytics is a generalisation obtained from __________ that can be used afterwords to generate __________ for new given __________. It can be seen as a __________ that can be used to make predictions. Thus, model induction is a __________ task.

   (e) **(10 points)** Develop an R function named as *findroot* that firstly requests coefficients of a quadratic function ($ax^2 + bx + c = 0$) from the user, then calculates Δ ($\Delta = b^2 - 4ac$) and finally computes the roots $x_{1,2}$ ($x_{1,2} = \frac{-b \pm \sqrt{\Delta}}{2a}$).

2. Write down the necessary R codes for the following questions.

   (a) **(3 points)** Install *nycflights13* package on RStudio and load *flights* dataset into the workspace.

```
TABLE I: tibble [336,776 x 19] (S3: tbl_df/tbl/data.frame)
 $ year          : int [1:336776] 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
 $ month         : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
 $ day           : int [1:336776] 1 1 1 1 1 1 1 1 1 1 ...
 $ dep_time      : int [1:336776] 517 533 542 544 554 554 555 557 557 558 ...
 $ sched_dep_time: int [1:336776] 515 529 540 545 600 558 600 600 600 600 ...
 $ dep_delay     : num [1:336776] 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
 $ arr_time      : int [1:336776] 830 850 923 1004 812 740 913 709 838 753 ...
 $ sched_arr_time: int [1:336776] 819 830 850 1022 837 728 854 723 846 745 ...
 $ arr_delay     : num [1:336776] 11 20 33 -18 -25 12 19 -14 -8 8 ...
 $ carrier       : chr [1:336776] "UA" "UA" "AA" "B6" ...
 $ flight        : int [1:336776] 1545 1714 1141 725 461 1696 507 5708 79 301 ...
 $ tailnum       : chr [1:336776] "N14228" "N24211" "N619AA" "N804JB" ...
 $ origin        : chr [1:336776] "EWR" "LGA" "JFK" "JFK" ...
 $ dest          : chr [1:336776] "IAH" "IAH" "MIA" "BQN" ...
 $ air_time      : num [1:336776] 227 227 160 183 116 150 158 53 140 138 ...
 $ distance      : num [1:336776] 1400 1416 1089 1576 762 ...
 $ hour          : num [1:336776] 5 5 5 5 6 5 6 6 6 6 ...
 $ minute        : num [1:336776] 15 29 40 45 0 58 0 0 0 0 ...
 $ time_hour     : POSIXct[1:336776], format: "2013-01-01␣05:00:00" ...
```

   (b) **(2 points)** Obtain Table I on the console of RStudio by using the *flights* data set.

   (c) **(10 points)** Install *tidyverse* package on RStudio, then summarise minimum, average, and maximum arrival delays with respect to a group of destinations in a separate manner.

```
TABLE II: JFK4JanMorning
   dep_time dep_delay arr_time arr_delay carrier dest  air_time distance
      <int>     <dbl>    <int>     <dbl> <chr>   <chr>    <dbl>    <dbl>
 1      604        -6      916        -5 B6      RSW        170     1074
 2      610        -5     1051        -9 B6      SJU        194     1598
 3      611        -4      807       -10 US      CLT         89      541
 4      615         0      851        -4 9E      ATL        118      760
 5      628        -2     1124       -16 AA      SJU        202     1598
 6      628        -2      947       -31 US      PHX        281     2153
 7      641         1      748        -1 B6      BOS         40      187
 8      643        -4      800       -10 B6      BUF         59      301
 9      650        -5      933       -57 DL      SLC        265     1990
10      653        -2      919        -2 B6      MSY        185     1182
11      655        -4      955        -4 AA      MCO        156      944
12      655         0      857       -45 B6      LAS        282     2248
13      700         0      941       -44 VX      LAX        318     2475
14      708        -7     1035        -5 UA      SFO        342     2586
15      712        -3     1022       -23 AA      MIA        156     1089
16      712        -3     1021       -14 AA      FLL        163     1069
17      715        15     1013       -21 DL      LAX        322     2475
18      716        16     1043        36 B6      FLL        174     1069
19      719        19     1102        17 DL      SFO        370     2586
20      721        21     1001       -13 B6      LAX        302     2475
21      721         0     1030        18 B6      MCO        149      944
22      721         6     1201        -5 B6      SJU        193     1598
23      727        -3     1025       -50 VX      SFO        328     2586
24      728        -2     1024       -36 AA      LAX        327     2475
25      732        -4      854         4 B6      SYR         47      209
26      733        -2      901         3 B6      ROC         58      264
27      735        -2     1030       -43 B6      SFO        319     2586
28      737        -3     1051       -14 DL      SEA        334     2422
29      744        -5     1057         3 B6      SRQ        159     1041
30      746         1     1106       -19 AA      SFO        343     2586
31      747        -3      940         0 9E      PIT         69      340
32      759         0     1120        -4 UA      SFO        346     2586
```

(d) **(15 points)** Transform the *flights* data set into *JFK4JanMorning* data set illustrated in Table II and save *JFK4JanMorning* data set onto the desktop of your computer in TSV format along with paying attention to the fact that *JFK4JanMorning* data set includes the information of flights originated from the JFK International Airport between the scheduled departure times from 06:00 to 08:00 (06:00 and 08:00 not included) on the 4th of January, 2013.

(e) **(5 points)** Reform the *JFK4JanMorning* data set by creating a new column named as *Velocity* (in km/h) with the help of *air_time* (in minutes) and *distance* (in miles) columns, and add the new column to the end of the other columns. (**Hint:** 1 statute mile equals to 1,609.344 metres.)

(f) **(5 points)** Reform the *JFK4JanMorning* data set by converting it from wide to long format via combining *dep_delay* and *arr_delay* columns under a new column named as *delay_type* and assign the values into the *delay_value*.

(g) **(10 points)** Plot the graph demonstrated in Figure 1 by utilising the *JFK4JanMorning* data set and assuming that you are en route to the San Francisco International Airport (SFO). Answer which carrier would be your choice for boarding and why?

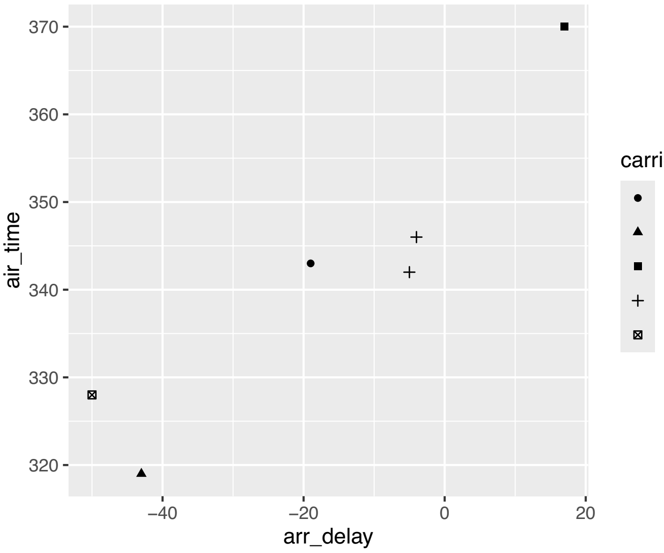

Figure 1: *air_time* vs *arr_delay* according to *carrier*

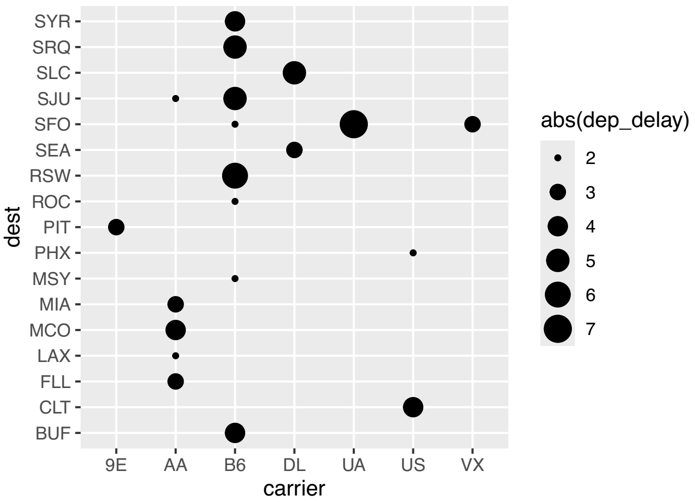

Figure 2: *dest* vs *carrier* according to the absolute value of *dep_delay*

(h) **(10 points)** Plot the graph indicated in Figure 2 by utilising the *JFK4JanMorning* data set for negative values of *dep_delay*. If you are en route to the San Francisco International Airport (SFO), answer which carrier would be your choice for boarding and why?