|  |
| --- |
| Instructions:   * You will have 2 hours to complete this test * You can use R or Python but scripts written in R will be given higher preference because this role will be working primarily in R * There are 10 questions * Label your script(s) with your name * Document and save your code in your script * Record your answers on the text boxes available on this word document * Upload your answers using this form: <https://forms.gle/aT1mHpXFTK7yFi9A6> |

1. Read the csv files (administrative\_site1.csv, administrative\_site2.csv and lab.csv) as **admin\_site1**, **admin\_site2** and **lab**, respectively. Create one data frame, **admin**, consisting of all rows in **admin\_site1** and **admin\_site2**.

(7 points)

1. Create a new *length\_of\_stay* variable as defined as discharge date and time minus admission date and time (in days). Calculate the mean *length\_of\_stay* for each hospital.

(13 points)

St. Joseph's Health Centre: 93.13

St. Michael's Hospital: 102.21

Mount Sinai Hospital: 99.97

Toronto Western Hospital: 100.89

Sunnybrook Health Sciences Centre: 97.61

1. Combine the **admin** and **lab** data using *ID* to create a complete dataset, retaining all IDs.

(5 points)

1. What similarities and differences do you see between the **admin\_site1** and **admin\_site2** data? How would you improve the format of the **admin** data?

(5 points)

Different date formats used.

One consistent format for all date columns.

## 5.

* 1. Transform the **lab** data from long format to wide format such that values in *test\_name* become columns (e.g., table below). If a test appears more than once for a subject (i.e., duplicated *test\_name* for the same *ID*), filter for the lowest *result\_value*.
  2. Name this data frame as lab\_wide and report column mean for “Urea plasma”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Bicarbonate plasma | Urea plasma | Creatinine Plasma | … | … |
| 1000 | 25.1 | 12.3 | NA |  |  |
| 1001 | 35.2 | NA | 548.2 |  |  |

(15 points)

Answer: 75.35869 mmol/L

## 6.

1. For **admin** data, remove any rows with missing values (NA) in 2 or more variables. Name the resulting data frame as **adm\_filtered**.
2. Report the missing rate (%) for each column in the **adm\_filtered** data.

(10 points)

ID: 0

admission\_date: 0

admission\_time: 0

discharge\_date: 0

discharge\_time: 0

hospital\_ID: 0

gender: 100

age: 100

disease\_status: 0

1. Make a copy of the **lab** data and name it **lab\_impute**. In **lab\_impute**, replace all *result\_value* collected in the year 2002 with NA. Then, impute these NAs with the mean value of the remaining records of the corresponding *test\_name* (e.g., replace the NAs of the *test\_name* “Bicarbonate plasma” with its mean value of non-NAs). Report the overall mean of *result\_value* after imputations for the *test\_name* “Bicarbonate plasma”.

(15 points)

Answer: Click or tap here to enter your answer.

1. For the **lab** data, find rows where the *provider\_name* contains “Jessica” or “Kelsey” (case insensitive). Compute and compare the mean *result\_value­* for the *test\_code* “NAPL” between providers containing “Jessica” vs. providers containing “Kelsey”.

(10 points)

Answer for “Jessica”: 131.681

Answer for “Kelsey”: 136

1. Using the **admin** data, create a function called “year\_output” to identify all rows with a specific year of hospital discharge and export the new data into a csv file. The function should have two arguments: year and output\_file\_name. Test your function with year = 2005 and output\_file\_name = Discharge\_2005.csv.

(10 points)

1. Create a function called “compute\_hospital\_mean” to compute the mean test *result\_value*. The function should have two arguments: *test\_code* and *hospital\_ID*. Merge the **admin** and **lab** data as appropriate or use previously created datasets. Return the result for *test\_code* = KPL and *hospital\_ID* = St. Michael’s Hospital.

(10 points)

Answer: 122.6063