|  |
| --- |
| Instructions:   * You will have 2.5 hours to complete this test. * You can use R or Python, but R is preferred as the role primarily uses R. * Page 4 is only for Data Scientist candidates. Page 5 is only for Data Engineer candidates. * There are 10 questions to be answered. * Record your answers either in a markdown format with your code, or in this document with code attached separately. * Name any scripts and documents you submit to start with “firstname.lastname” * Upload your answers using this form: <https://forms.gle/aT1mHpXFTK7yFi9A6> |

* 1. Read in the csv files (admissions\_surg.csv, admissions\_med.csv and img.csv) as **admissions\_surg**, **admissions\_med**, and **imaging** data frames.   
     Briefly describe similarities and differences you see between these datasets. How would you improve the overall format of this data?

Admissions\_surg and admissions\_med has same columns but in different format. The primary key for all three dataframes are ‘ID’. I rename the columns in admissions\_surg so that it’s the same as admissions\_med.

The date and time columns are converted to date time datatype.

* 1. Create one data frame called **admissions\_img**, consisting of all rows in **admissions\_surg** and **admissions\_med,** merged with the **imaging** data using *ID* (retaining all IDs from both).

1. In **admissions\_img**, create a new *length\_of\_stay* variable defined as discharge date and time minus admission date and time (in days). Calculate the mean *length\_of\_stay* for each department. Briefly describe how you dealt with missing data.

Addiction Services: 101

General Internal Medicine: 98

Oncology: 103

Palliative Care: 96.

General Surgery: 101

Obstetrics: 105

Missing data were ignored when calculating the mean.

* 1. In **imaging**, filter to the first performed test for each *test\_name*. Save this data frame as **q3\_df**.
  2. Transform **q3\_df** into wide format such that each *test\_name* becomes a column displaying the *performed\_date* of that test (see example table below). Display the head of the table.

Example table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **US ABDOMEN** | **Doppler Ultrasound** | **NECK AND HEAD CT** | **…** |
| 95964 | 1980-01-13 | NA | NA | … |
| 75459 | NA | 1980-01-30 | NA | … |
| 36730 | NA | NA | 1980-02-04 | … |

1. From **admissions\_img,** remove any rows with missing values (NA) in 2 or more variables and name the resulting data frame **q4\_df**. Report any non-zero missing rate (%) in this data frame.

Non-zero missing rate(%):

ordered\_date\_time 33.96

discharge\_time 11.48

age 5.98

main\_diagnosis\_name 8.18

1. In **admissions\_img,** impute the missing *age*, *test\_name,* and *technician\_name* values in any way you see fit and name the resulting data frame **q5\_df**. Briefly describe why you chose that method and print the mean of *age* after imputation.

I filled missing age values using the average age 56.87. Since the missing rate of age is about 6%. It is relatively small. So fill it with mean won’t have a great impact on the feature distribution.

There are no missing value in test\_name and technician\_name.

1. Using the **imaging** data, create a frequency table of unique *test\_name*, with decreasing order of frequency. Name this table **img\_mapper**.
2. In **img\_mapper**, add a column named *mapped\_test* which classifies each unique *test\_name* into 3 categories: "US" (or Ultrasound), "CT" and "Others"
3. In **img\_mapper,** add another column named *mapped\_bodypart*, which classifies each unique *test\_name* into 3 categories, "Neck/Head", "Abdomen/pelvis", "Others". Display the table head.

Example table:

|  |  |  |  |
| --- | --- | --- | --- |
| **raw\_test\_name** | **frequency** | **mapped\_test** | **mapped\_bodypart** |
| ct neck and head | 408 | CT | Neck/Head |
| US ABDOMEN | 396 | US | Abdomen/Pelvis |
| Doppler Ultrasound | 233 | US | Others |
| … | … | … | … |

1. Using **img\_mapper**, create a function which takes **imaging** data, **img\_mapper,** and year as input to compute the frequencies of each pair of *mapped\_test* and *mapped\_bodypart* performed for that year (see sample table below).

Example table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Abdomen/Pelvis** | **Neck/Head** | **Others** |
| **CT** | ? | ? | ? |
| **US** | ? | ? | ? |
| **Others** | ? | ? | ? |

Report the number of ultrasound tests were performed on Neck/Head in 2020. What would you consider to improve this task in terms of efficiency (e.g. in case for + 200 different test names)?

Click or tap here to enter your answer.

**Questions 8-DS, 9-DS, and 10-DS are only to be answered by Data Scientist candidates**

**Data Engineer candidates go to page 5.**

8-DS. Using **admissions\_img**, Create a clearly labeled plot to demonstrate the relationship between *department* and *length\_of\_stay*. Briefly explain why you chose this plot type.

I choose to plot it in box plot to display the distribution of length\_of\_stay by department.

## 

## 9-DS. You are interested in predicting *length\_of\_stay* and exploring which variables have the largest effect on it.

## Fit a linear regression model using *length\_of\_stay* as the dependent variable and *age*, *gender*, and *department* as the independent variables. Interpret the results.

## Describe other statistical models would you consider and why. What other variables would you be interested in adding to the model (they don’t have to be in the provided datasets)? How would you evaluate the model’s accuracy?

Coefficients:

[ 1.83942789 -3.35255598 1.26845565 2.89998791 0.61761521 -3.27293068

0.00546781 2.14006798]

Mean squared error: 3043.91

Coefficient of determination: -0.00

## 10-DS. You are interested in investigating whether *brief\_report* either will or will not contain the words “cancer” or “tumor”. Create a variable that indicates if a row contains one of these words, then fit a statistical model of your choosing to it. Justify your choices and evaluate the accuracy of your model. How would you improve it if you had to iterate on it?

Click or tap here to enter your answer.

**Questions 8-DE, 9-DE, and 10-DE are only to be answered by Data Engineer candidates**

Consider a database with the **room\_transfers** table that has the following fields and datatypes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **patient\_id** | **admission\_id** | **enter\_datetime** | **exit\_datetime** | **room** |
| (UUID) | (UUID) | (datetime) | (datetime) | (“A”, “B”, or “C”) |

This table records the time patients are in specific hospital rooms. A patient can move between multiple rooms in a single hospital admission. A patient can have multiple hospital admissions.

8-DE. Write a SQL query to find the room(s) with the highest average stay time (in days).

Click or tap here to enter your answer.

9-DE. Write a SQL query to find the average amount of time elapsed (in hours) between admissions over all patients.

Click or tap here to enter your answer.

10-DE. Write a SQL query to find the number of patients who left room “A” for room “B” then went back to room “A” in a single admission all within an hour.

Click or tap here to enter your answer.