**Part A**

**Research Question:**

*Is there a relationship between the age of the patient and whether or not they are readmitted?*

The organization can benefit from analyzing this data because it can help predict which patients are more or less likely to be readmitted. Based on these predictions, they can make changes to their care to patients who may be at high risk of readmission, which, in this case, would be the age of the patient. Ultimately, they can be more prepared, avoid penalties for readmissions, and have improved post-discharge health of patients.

The two variables that we would focus on in order to answer are research question are represented by the “ReAdmis” and “Age” columns. ReAdmis is categorical and qualitative in nature, denoting whether or not the patient was readmitted to the hospital (indicated by a yes or no). Age is numeric and continuous, and it refers to the age (in years) of the patient

**Part B**

A picture containing diagram

Description automatically generated

**Results of t-test:**

* t-value = 1.5811
* p-value = 0.1139
* A p-value of 0.1139 is greater than our alpha p-value of 0.05

*This indicates no statistical significance and no relationship between age and readmission status.*

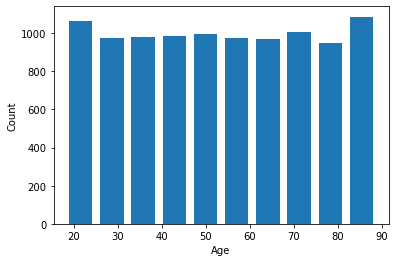
I chose to run a t-test because it allows us to test a hypothesis with one numeric variable (in this scenario, age) and one categorical variable (particularly a binary variable, which in this scenario is readmission). The question I asked in part A1 had one numerical variable and one categorical, binary variable.

See attached Jupyter notebook for t-test code.

**Part C**

**Univariate Exploration:**

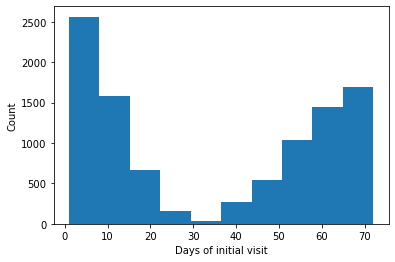
***Continuous variable 1: Age***



|  |  |
| --- | --- |
| Count | 10000 |
| Mean | 53.5117 |
| StdDev | 20.638538 |
| Min | 18 |
| 25% | 36 |
| 50% | 53 |
| 75% | 71 |
| Max | 89 |

The distribution of the age variable appears to be a uniform distribution, with all ages being relatively similar between 18 and 89 years. Various statistics about the ages, such as mean, standard deviation, and median are in the table above.

***Continuous variable 2: Initial\_days***



|  |  |
| --- | --- |
| Count | 10000 |
| Mean | 34.455299 |
| StdDev | 26.309341 |
| Min | 1.001981 |
| 25% | 7.896215 |
| 50% | 35.836244 |
| 75% | 61.16102 |
| Max | 71.98149 |

The distribution of the initial days variable appears to be bimodal, with two distinct peaks at the minimum and maximum values. Various statistics about the variable, such as mean, standard deviation, and median are in the table above.

***Categorical variable 1: ReAdmisChart

Description automatically generated***

Chart, bar chart

Description automatically generated***Categorical variable 2: Services***

**Part D**

**Bivariate Exploration:**

***Continuous variables: TotalCharge and Initial\_days***

A picture containing text, sky, smoke, outdoor

Description automatically generated

There appears to be a strong positive relationship between total charge and initial days. This makes sense because most hospitals charge per day, so the longer a patient stays, the more they will be charged.

***Continuous variables: VitD\_levels and Age***

***A picture containing background pattern

Description automatically generated***

There appears to be no relationship between age and Vitamin D levels of patients. Vitamin D levels appear to have a consistent approximate mean throughout all ages. This is somewhat surprising, as I would think older patients would be more vitamin-deficient – I would have predicted a strong negative relationship between age and Vitamin D levels.

Chart, box and whisker chart

Description automatically generated***Categorical variables: ReAdmis and Initial\_days***

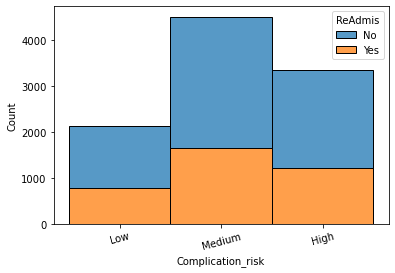
Chart, histogram

Description automatically generated

These visualizations show the difference in days of initial visit between patients who are readmitted and those who are not. The median number of days for readmitted patients is approximately 65, while the median for non-readmitted patients is approximately 10.

\*Note: one of these variables is categorical and one is numeric/continuous.

***Categorical variables: ReAdmis and Complication\_risk***



|  |  |  |
| --- | --- | --- |
| **Readmitted** | **Count** | **100%** |
| Low | 782 | 21.3% |
| Medium | 1664 | 45.4% |
| High | 1223 | 33.3% |
| **Not Readmitted** |  | **100%** |
| Low | 1343 | 21.2% |
| Medium | 2853 | 45.1% |
| High | 2135 | 33.7% |

This visualization shows the difference in complication risk in patients who are readmitted and those who are not. The table on the left is grouped by readmission status, while the histogram on the right is grouped by complication risk. These visualizations have a few surprising implications, including the fact that the proportions of low, medium, or high-risk patients (indicated by the percentages in the table) are relatively similar between both readmission status. This implies that a patient is approximately equally likely to be readmitted, independent of their deemed complication risk. In short, it doesn’t matter what your complication risk is - your chances of being readmitted are about 36%.

**Part E**

Null (H0): There is no relationship between a patient’s age and whether or not they are readmitted

Hypothesis (H1): There is a relationship between a patient’s age and whether or not they are readmitted.

H0: μreadmit – μnot = 0

H1: μreadmit - μnot ≠ 0

where:

μreadmit and μnot are the population mean ages for readmitted patients and not readmitted patients, respectively.

The p-value of 0.1139 is greater than our alpha p-value of 0.05

This indicates no statistical significance and no relationship between age and readmission status.

This data analysis has various significant limitations, including, but not limited to:

- an unknown population size (and possibly too small a sample size)

- unverified accuracy of data

- unknown randomness of data

- non-normal distribution of age variable (instead a uniform distribution) for t-test

- potential to incorrectly implicate correlation with causation

I would make a few recommendations, starting with collecting more data about patients who get readmitted, including reason for readmission, post-discharge care, any medications used, and a more specific reason for admission. I would also recommend looking further into the correlation between readmission rates and initial days of hospital visit. We found a clear and extreme correlation, to the extent that almost all patients whose initial visit was over 50 days were readmitted, and almost all patients with initial visits under 50 days were not readmitted. If that is accurate, it’s clear who will be readmitted, and we need to adjust our care accordingly.