D207 Performance Assessment

OEM2 - OEM2 Task 1: EDA - Exploratory Data Analysis

Exploratory Data Analysis - D207  
PRFA - OEM2

Table of Contents

[A.  Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following: 1](#_Toc126222044)

[B.  Describe the data analysis by doing the following: 3](#_Toc126222045)

[C.  Identify the distribution of two continuous variables and two categorical variables using univariate statistics from your cleaned and prepared data. 6](#_Toc126222046)

[D.  Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data. 8](#_Toc126222047)

[E.  Summarize the implications of your data analysis by doing the following: 10](#_Toc126222048)

[F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used. 11](#_Toc126222049)

[G.  Reference the web sources used to acquire segments of third-party code to support the analysis. 11](#_Toc126222050)

[H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized. 11](#_Toc126222051)

[I.   Demonstrate professional communication in the content and presentation of your submission. 12](#_Toc126222052)

# A.  Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:

1. Provide **one** question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).  
     
    We will be using the medical dataset. The organization is going to use Python and its data-science libraries to conduct exploratory analytics to identify the related variables to make better decisions in the future.  
     
   The hospital wants to know if Initial Admission Type (Initial\_admin) and Readmissions (ReAdmis) are dependent or independent variables? To find the answer for the chi-square test will be used.
2. Explain how stakeholders in the organization could benefit from an analysis of the data.  
     
    “Every year, the Centers for Medicare & Medicaid Services (CMS) penalizes hospitals with excessively high readmission rates” (Allen, 2023). Finding the variables that lead to higher readmission rates will help the hospital avoid the penalty and provide better healthcare overall.
3. Identify *all* of the data in your data set that are relevant to answering your question in part A1.  
     
   Western Governors University provided the dataset ‘medical\_clean.csv’ and data dictionary ‘D207 D208 D209 Medical Data Considerations and Dictionary.pdf’ (Western Governors University [WGU], n.d.) describing all of the variables collection for the analysis of readmissions.  
     
   The ‘medical\_clean.csv’ dataset was reviewed to assist in being familiar with the data before the analyzing process began.   
     
   To assist in finding relevant variables for this exploratory data analysis we prepared the data in a few different ways. First, we checked the data for blanks and NAs. Next, we made sure to convert the ‘chr’ and ‘str’ data types to categorical. Then we removed categorical variables that had a high cardinality based on formula: Cardinality > 3 to 5 levels.  
     
   The following variables are being retained for analysis:  
     
   ReAdmis, Initial\_admin which are the categoricals that we will be running the chi-square analysis on. Then we’ll also retain Income, VitD\_levels which we will uses for B and C for our univariate and bivariate statistics.  
   Summary table of retained variables;

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Data Type | Description | Example data |
| ReAdmis | Categorical (Bi-Nomial) | Readmitted within a month of release | Yes/No |
| Initial\_admin | Categorical (Nominal) | The means in which the patient was admitted into the hospital. | Elective Admission  Observation Admission  Emergency Admission |
| Income | Numerical (Continuous) | Age of the patient | 14370.14  39741.49  1209.56 |
| VitD\_levels | Numerical (Continuous) | Patients Vitamin D level. | 19.141466  18.940352  18.057507  16.576858 |

# B.  Describe the data analysis by doing the following:

1.  Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:

•   chi-square

•   t-test

•   ANOVA

#Contingency Table  
cross\_tab = pd.crosstab(medical\_data\_prepared['Initial\_admin']  
 ,medical\_data\_prepared['ReAdmis'])   
cross\_tab

#Observed values are Yes and No per Initial\_admin.  
observed\_values = cross\_tab.values  
print('Observed Values: \n', observed\_values)

#Chi-square test of independence of variables in a contingency table.  
test\_statistic, p\_value, degree\_of\_freedom, expected\_frequency = chi2\_contingency(observed\_values)

probability = .95  
critical\_value = chi2.ppf(probability, degree\_of\_freedom)  
  
print(f'Probability = {probability}, Chi-Square Statistic = {chi\_square\_statistic}, Critical Value = {critical\_value}')  
  
if critical\_value <= abs(chi\_square\_statistic):  
 print('Dependent (Reject Null Hypothesis [H0])')  
else:  
 print('Independent (Fail to reject Null Hypothesis [0])')

print(f'Degrees Of Freedom = {degree\_of\_freedom}')  
print(f'Expected Frequency = {expected\_frequency}')

probability = .95  
critical\_value = chi2.ppf(probability, degree\_of\_freedom)  
  
print(f'Probability = {probability}, Chi-Square Statistic = {chi\_square\_statistic}, Critical Value = {critical\_value}')  
  
if critical\_value <= abs(chi\_square\_statistic):  
 print('Reject Null Hypothesis (H0), variables are dependent.')  
else:  
 print('Fail to reject Null Hypothesis (H0), variables are independent.')

significance\_level\_alpha = 1.0 - probability  
  
print(f'Significance = {significance\_level\_alpha}, Probability = {probability}')  
  
if probability <= significance\_level\_alpha:  
 print('Reject Null Hypothesis (H0), variables are dependent.')  
else:  
 print('Fail to reject Null Hypothesis (H0), variables are independent.')

2.  Provide the output and the results of *any* calculations from the analysis you performed

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

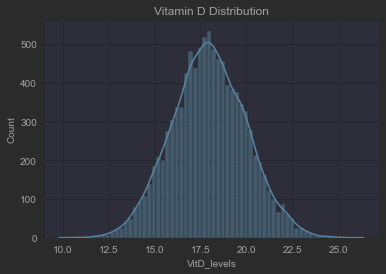
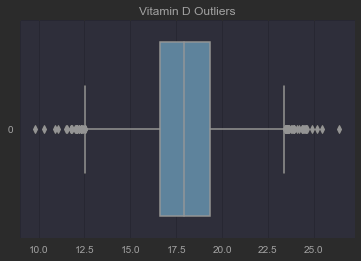




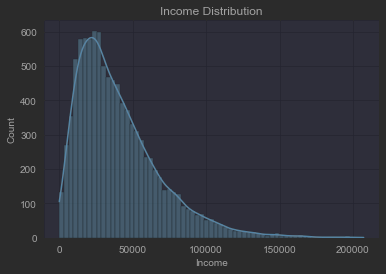
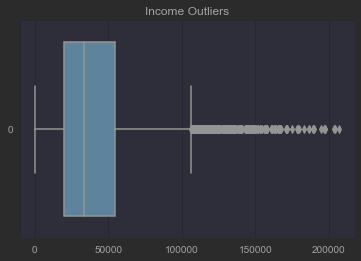
3.  Justify why you chose this analysis technique.

First it should be noted that we cannot use t-Test or ANOVA test because the values of the hypothesis are categorical values.   
  
Secondly, the type of Chi-Square test is the ‘Test of independence’ version as the hospital wants to know if Initial Admission Type affects readmission rates, we wanted to use a test against a Null Hypothesis. That null hypothesis is as follows, H0: The factors are independent.   
  
If we fail to reject H0 then we know that these variables are independent and not much of a factor in readmission. If H0 is rejected, then we know that the two variables are dependent on one another, and the hospital should focus on Initial Admission Types to help reduce readmission rates to provide better care and avoid penalties.

# C.  Identify the distribution of **two** continuous variables and **two** categorical variables using univariate statistics from your cleaned and prepared data.

1. Represent your findings in Part C, visually as part of your submission.  
     
   VitD\_level  
     
    

Income

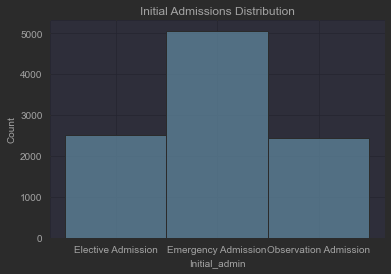
 

Readmission



ReAdmis  
No 6331  
Yes 3669  
dtype: int64

Initial\_admin



Initial\_admin  
Elective Admission 2504  
Emergency Admission 5060  
Observation Admission 2436  
dtype: int64

*Note: To draw a graph or visualization, you may use one or a combination of the following:*

*- A spreadsheet program, such as Excel (\*.xls)*

*- A graphics program, such as Paint (\*.jpeg, \*.gif)*

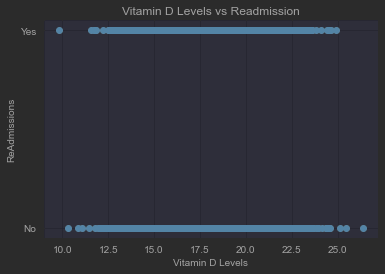
*- A word-processing program, such as Word (\*.rtf)*

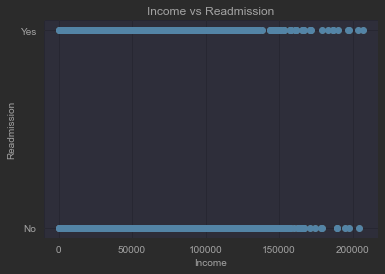
*- A scanned hand-drawn graph (\*.jpeg, \*.gif)*

# D.  Identify the distribution of **two** continuous variables and **two** categorical variables using bivariate statistics from your cleaned and prepared data.

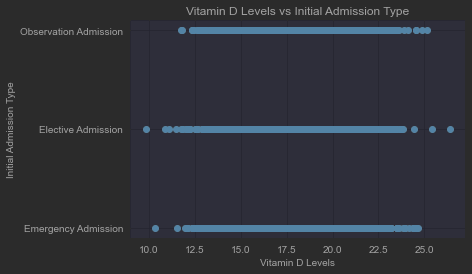
1. Represent your findings in Part D, visually as part of your submission.

Vitamin D Levels (x) and Readmission Distribution (y)

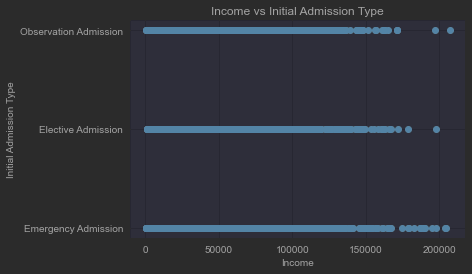


Income (x) and Readmission Distribution (y)  
  


Vitamin D Levels (x) and Initial Admission Type (y)



Income (x) and Initial Admission Type



*Note: To draw a graph or visualization, you may use one or a combination of the following:*

*- A spreadsheet program, such as Excel (\*.xls)*

*- A graphics program, such as Paint (\*.jpeg, \*.gif)*

*- A word-processing program, such as Word (\*.rtf)*

*- A scanned hand-drawn graph (\*.jpeg, \*.gif)*

# E.  Summarize the implications of your data analysis by doing the following:

1.  Discuss the results of the hypothesis test.  
  
Based on the results of the Chi-Square test we failed to reject the null hypothesis, which is H0: The factors are independent (Initial Admission Type and Readmission are not related). Because we failed to reject, we know that Initial Admission Types are independent of Readmission. The hospital can investigate other variables within their dataset that affect readmissions.

2.  Discuss the limitations of your data analysis.  
  
Limitations of this analysis can be found in that the question was extremely specific given the amount of available data in the medical dataset. The hospitals question could be expanded in a way that more of the suggested tests t-test, ANOVA, etc can be used to start isolating which variables do in fact affect readmission.

3.  Recommend a course of action based on your results.

Based on the rejection of the null hypothesis the recommended course of action would be to continue further analysis on additional variables. If they are numerical, we can add t-Test and ANOVA into the mix. We should also run some more analysis tests on groups of variables that affect readmission. We would need to be mindful of the Type I Error of running to many tests.

# F.  Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3f9ce88f-94d9-4661-a117-af9d010e4e27>

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.*

*To submit your recording, upload it to the Panopto drop box titled “Exploratory Data Analysis – OEM2 \ D207.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

# G.  Reference the web sources used to acquire segments of third-party code to support the analysis.

  N/A

# H.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

* Allen, J. (2023, January 4). Understanding The 2023 Medicare Hospital Readmission Penalty. What I’ve Learned as a Hospital Medical Director. Retrieved January 31, 2023, from <https://hospitalmedicaldirector.com/understanding-the-2023-medicare-hospital-readmission-penalty/>
* Western Governors University [WGU]. (n.d.). D207 Exploratory Data Analysis | D208 Predictive Modeling | D209 Data Mining I. WGU. Retrieved January 31, 2023, from <https://access.wgu.edu/ASP3/aap/content/g9rke9s0rlc9ejd92md0.html>

# I.   Demonstrate professional communication in the content and presentation of your submission.