Insightful Data Exploration

Grace Ellen Steinke

MSDA Program, Western Governor's University

D207: Exploratory Data Analysis

December 27, 2021

A1: Question for Analysis

Are hospital readmission rates influenced by residential area type?

H1: The number of hospital readmissions differ between rural, suburban, and urban residential area types.

H0: The number of hospital readmissions do not differ between rural, suburban, and urban residential area types.

(Nigam, 2021) (DataCamp, 2021)

In other words, I want to explore the possibility that whether or not someone was readmitted to the hospital is dependent on what type of area they live in. To do this, we can take a look at our sample data and see that yes, the hospital readmission rates do change across residential area types. Now we must determine if that is simply due to chance, or if it is statistically representative of the larger population – therefore providing insights into possible trends surrounding patient readmission rates. Our findings could present a specific area in which hospital readmission rates are statistically more prevalent than others.

A2: Benefit From Analysis

This information can help stakeholders in the organization to better understand how to potentially reduce hospital readmission rates. Doing this presents an opportunity to lower health care costs, improve quality of care, and increase patient satisfaction all at the same time.

A3: Data Identification

Including 10,000 patients and 50 variables, the data set is described in-depth below:

Variable Name	Data Type	Variable Description	Example
CaseOrder	int64	Placeholder variable to preserve original order of the raw data file	1
Customer_id	object	Unique patient ID	C412403
Interaction	object	Unique ID related to patient transactions, procedures, and admissions	8cd49b13-f45a-4b47-a2bd-173ffa932c2f
UID	object	Unique ID related to patient transactions, procedures, and admissions	3a83ddb66e2ae73798bdf1d705dc0932
City	object	Patient city of residence as listed on the billing statement	Eva
State	object	Patient state of residence as listed on the billing statement	AL
County	object	Patient county of residence as listed on the billing statement	Morgan

Zip	int64	Patient zip code of residence as listed on the billing statement	35621
Lng	float64	GPS coordinate (longitude) of patience residence as listed on the billing statement	-86.7251
Lat	float64	GPS coordinate (latitude) of patience residence as listed on the billing statement	34.3496
Population	int64	Population within a mile radius of patient, based on census data	2951
Area	object	Area type (rural, urban, suburban), based on unofficial census data	Suburban
TimeZone	object	Time zone of patient residence based on patient's sign-up information	America/Chicago
Job	object	Job of the patient (or primary insurance holder) as reported in the admissions information	Psychologist, sport and exercise
Children	float64	Number of children in the patient's household as reported in the admissions information	1

Age	float64	Age of the patient as reported in admissions information	53
Education	object	Highest earned degree of patient as reported in admissions information	Some College, Less than 1 Year
Employment	object	Employment status of patient as reported in admissions information	Full Time
Income	float64	Annual income of the patient (or primary insurance holder) as reported at time of admission	86575.93
Marital	object	Marital status of the patient (or primary insurance holder) as reported on admissions information	Divorced
Gender	object	Customer self-identification as male, female, or nonbinary	Male
ReAdmis	object	Whether the patient was readmitted within a month of release or not (yes, no)	No
VitD_levels	float64	The patient's vitamin D levels as measured in ng/mL	17.80233

Doc_visits	int64	Number of times the primary physician visited the patient during the initial hospitalization	6
Full_meals_eaten	int64	Number of full meals the patient ate while hospitalized (partial meals count as 0, and some patients had more than three meals in a day if requested)	0
VitD_supp	int64	The number of times that vitamin D supplements were administered to the patient	0
Soft_drink	object	Whether the patient habitually drinks three or more sodas in a day (yes, no)	No
Initial_admin	object	The means by which the patient was admitted into the hospital initially (emergency admission, elective admission, observation)	Emergency Admission
HighBlood	object	Whether the patient has high blood pressure (yes, no)	Yes
Stroke	object	Whether the patient has had a stroke (yes, no)	No
Complication_risk	object	Level of complication risk for the patient as assessed by a primary patient assessment (high, medium, low)	Medium

Overweight	float64	Whether the patient is considered overweight based on age, gender, and height (1 = yes; 0 = no))	0
Arthritis	object	Whether the patient has arthritis (yes, no)	Yes
Diabetes	object	Whether the patient has diabetes (yes, no)	Yes
Hyperlipidemia	object	Whether the patient has hyperlipidemia (yes, no)	No
BackPain	object	Whether the patient has chronic back pain (yes, no)	Yes
Anxiety	float64	Whether the patient has an anxiety disorder (1 = yes, 0 = no)	1
Allergic_rhinitis	object	Whether the patient has allergic rhinitis (yes, no)	Yes
Reflux_esophagitis	object	Whether the patient has reflux esophagitis (yes, no)	No

Asthma	object	Whether the patient has asthma (yes, no)	Yes
Services	object	Primary service the patient received while hospitalized (blook work, intravenous, CT scan, MRI)	Blood Work
Initial_days	float64	The number of days the patient stayed in the hospital during the initial visit	10.58577
TotalCharge	float64	The amount charged to the patient daily. This value reflects an average per patient based on the total charge divided by the number of days hospitalized. This amount reflects the typical charges billed to patients, not including specialized treatments.	3191.049
Additional_charges	float64	The average amount charged to the patient for miscellaneous procedures, treatments, medicines, anesthesiology, etc.	17939.4
Item1	int64	Response to question 1 of an 8-question survey asking customers to rate the importance of timely admission (1 = most important; 8 = least important)	3
Item2	int64	Response to question 2 of an 8-question survey asking customers to rate the importance of timely treatment (1 = most important; 8 = least important)	3
Item3	int64	Response to question 3 of an 8-question survey asking customers to rate the importance of timely visits (1 = most important; 8 = least important)	2

Item4	int64	Response to question 4 of an 8-question survey asking customers to rate the importance of reliability (1 = most important; 8 = least important)	2
Item5	int64	Response to question 5 of an 8-question survey asking customers to rate the importance of options (1 = most important; 8 = least important)	4
Item6	int64	Response to question 6 of an 8-question survey asking customers to rate the importance of hours of treatment (1 = most important; 8 = least important)	3
Item7	int64	Response to question 7 of an 8-question survey asking customers to rate the importance of courteous staff (1 = most important; 8 = least important)	3
Item8	int64	Response to question 8 of an 8-question survey asking customers to rate the importance of evidence of active listening from doctor (1 = most important; 8 = least important)	4
index	int64	Python-friendly index field created to encourage greater ease and organization when manipulating the data set. Each observation is assigned a numeric index value (0-9999)	0

B2: Output & B3: Justification

I chose to use the chi-squared hypothesis testing technique because it is well-known for helping us to better understand and interpret the relationship (if any) between two categorical variables (2021). Since we are examining the variables ‘ReAdmis’ and ‘Area’ from the medical data set, this test allows us to see whether or not these variables are independent from one another. Prior to using the chi-squared test for independence to calculate the p-value, it is important to determine the cutoff point for rejecting the null hypothesis. For the purposes of this study, alpha is set to 0.05:

```
#Significance level
alpha = 0.05
```

(DataCamp, 2021)

This provides us with a mere 5% risk of telling us that a difference exists when there is no actual difference present. Hence, providing stronger evidence towards the conclusion – resulting in greater integrity and credibility.

```
#Calculation of Chisquare test statistic
chi_square = 0
rows = med_df['Area'].unique()
columns = med_df['ReAdmis'].unique()
for i in columns:
    for j in rows:
        O = data_crosstab[i][j]
        E = data_crosstab[i]['Total'] * data_crosstab['Total'][j] / data_crosstab['Total']['Total']
        chi_square += (O-E)**2/E
```

(Nigam, 2021)

```
#The p-value approach

print("The p-value approach to hypothesis testing in the decision rule")

p_value = 1 - stats.norm.cdf(chi_square, (len(columns)-1))

conclusion = "Failed to reject the null hypothesis."

if p_value <= alpha:

    conclusion = "Null Hypothesis is rejected."



print("chisquare-score is:", chi_square, " and p value is:", p_value)

print(conclusion)
```

(Nigam, 2021)

C: Univariate statistics

The following code was used to calculate the ‘five-number summary’ for the ‘Age’ column of the medical readmission data set:

```
#Calculate the quartiles of Age

print(np.quantile(med_df['Age'], [0, 0.25, 0.5, 0.75, 1]))
```

(DataCamp, 2021)

The output is shown below:

```
[18. 36. 53. 71. 89.]
```

The following code was used to calculate the ‘five-number summary’ for the ‘Income’ column of the medical readmission data set:

#Calculate the quartiles of Income

print(np.quantile(med_df['Income'], [0, 0.25, 0.5, 0.75, 1]))

(DataCamp, 2021)

The output is shown below:

```
[1.54080000e+02 1.95987750e+04 3.37684200e+04 5.42964025e+04
 2.07249100e+05]
```

The following code was used to create a frequency table to count how many of each marital status are included in the data set:

```
#Find frequency of each marital status

pd.crosstab(index=med_df['Marital'], columns='count')
```

(Lumen Learning, 2013)

The output is shown below:

col_0	count
Marital	
Divorced	1961
Married	2023
Never Married	1984
Separated	1987
Widowed	2045

The following code was used to create a frequency table to count how many of each gender are included in the data set:

#Find frequency of each gender

pd.crosstab(index=med_df['Gender'], columns='count')

(Lumen Learning, 2013)

The output is shown below:

Gender	col_0 count
Female	5018
Male	4768
Nonbinary	214

C1: Visual of findings

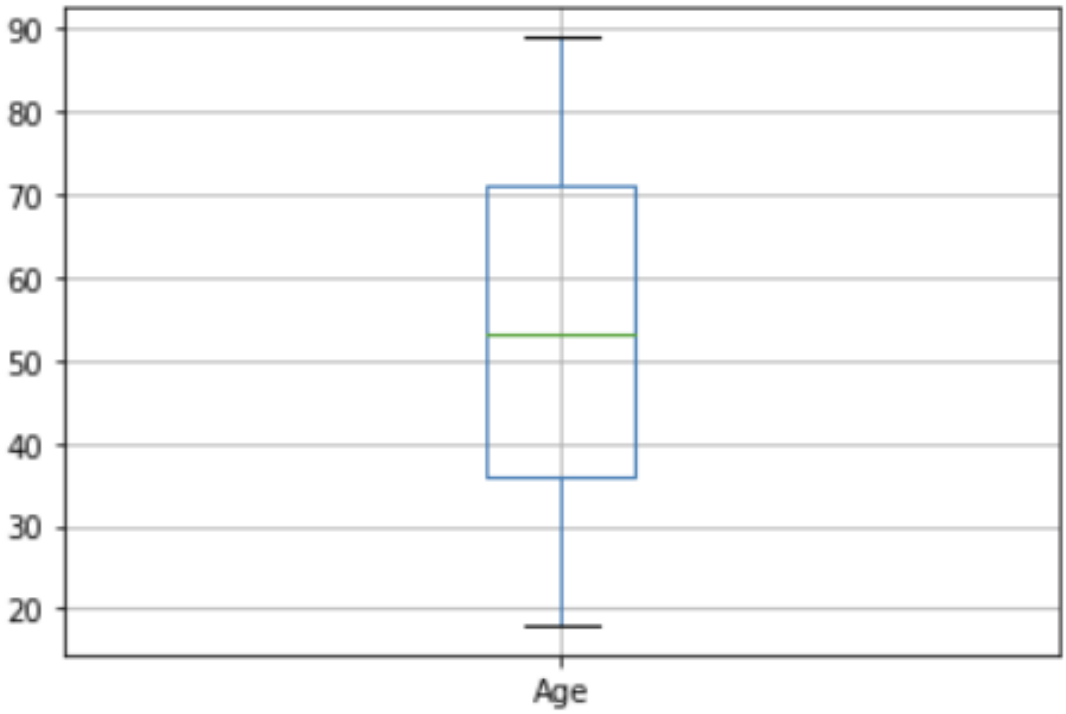
The following code was used to create a box plot which provides a visual representation of the 'five-number summary' for the 'Age' column of the medical readmission data set:

#Create Box Plot to visually represent the distribution of Age

med_df.boxplot(column = ['Age'])

(Lumen Learning, 2013)

The output is shown below:

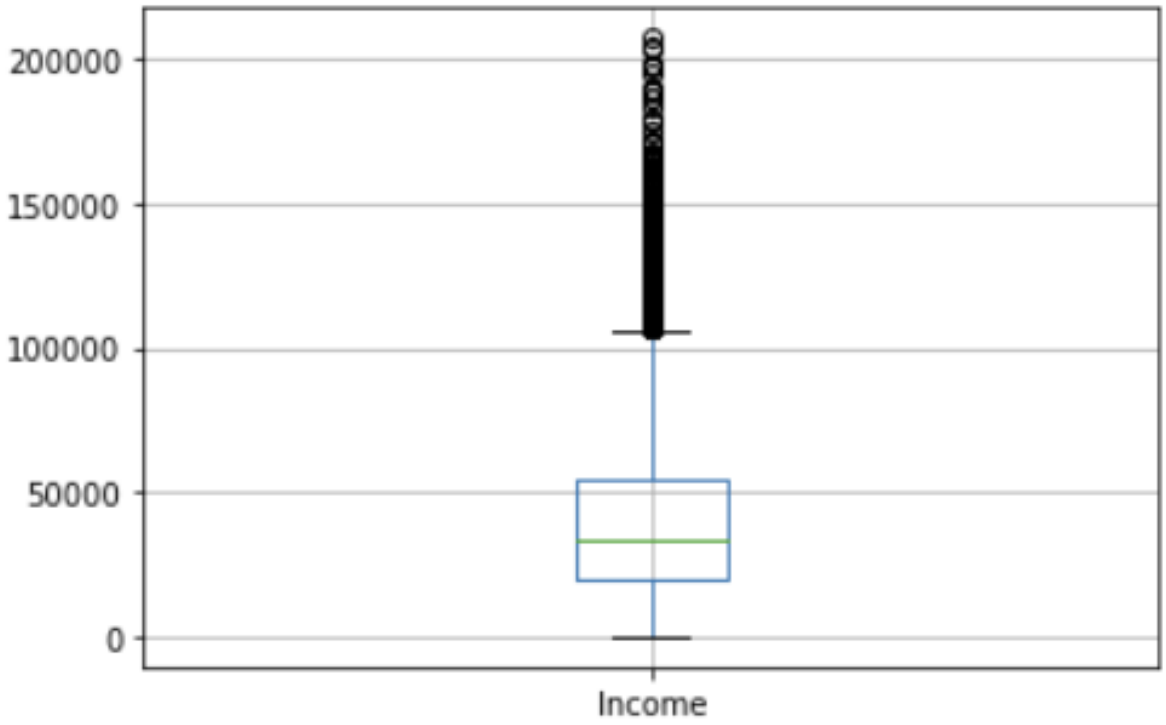


The following code was used to create a box plot which provides a visual representation of the ‘five-number summary for the ‘Income’ column of the medical readmission data set:

```
#Create Box Plot to visually represent the distribution of Income

med_df.boxplot(column = ['Income'])
```

(Lumen Learning, 2013)

The output is shown below:



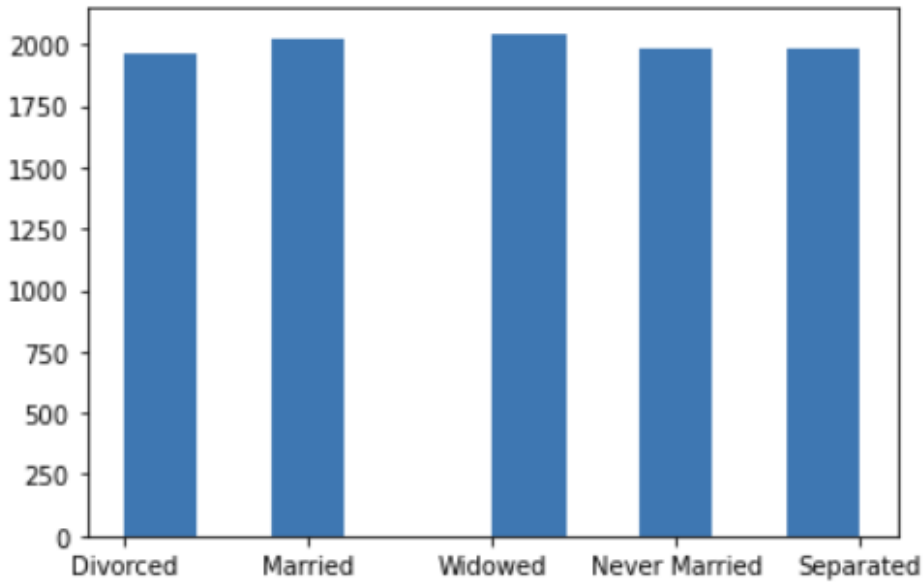
The following code was used to create a histogram which provides a visual representation of a frequency distribution for each category of the ‘Marital’ column of the medical readmission data set:

```
#Create histogram to show frequency distribution of each marital status

plt.hist(med_df['Marital'])
```

plt.show()

(Lumen Learning, 2013)

The output is shown below:



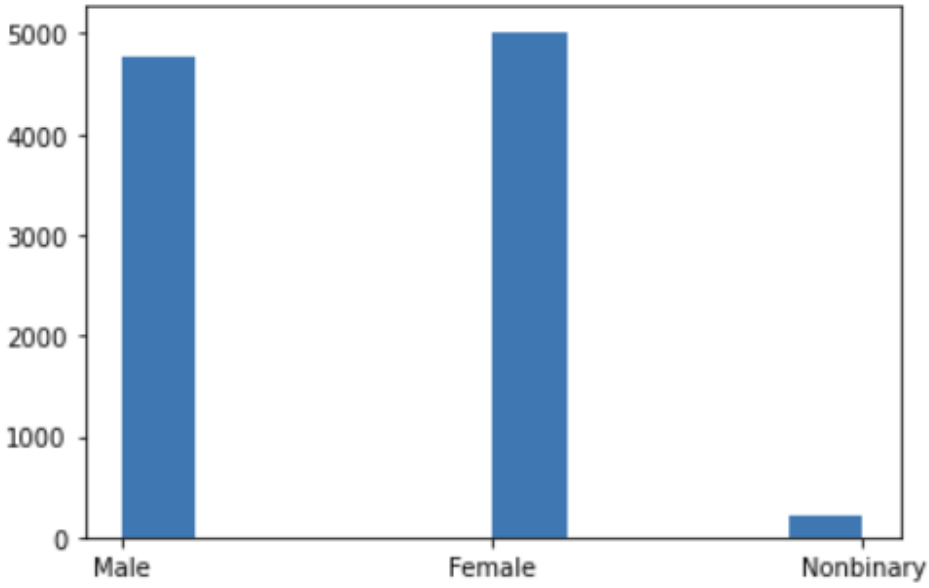
The following code was used to create a histogram which provides a visual representation of a frequency distribution for each category of the ‘Gender’ column of the medical readmission data set:

#Find frequency of each gender

pd.crosstab(index=med_df['Gender'], columns='count')

(Lumen Learning, 2013)

The output is shown below:



D: Bivariate statistics

The following code is used to calculate the variance and standard deviation of the ‘Income’ variable for each category/status listed in the ‘Marital’ column:

#Print variance and sd of Income for each Marital status

print(med_df.groupby('Marital')['Income'].agg([np.var, np.std]))

(2021)

The output is shown below:

Marital	var	std
Divorced	8.750829e+08	29581.799691
Married	8.171095e+08	28585.127379
Never Married	8.182316e+08	28604.747602
Separated	7.826233e+08	27975.404799
Widowed	7.758870e+08	27854.747824

The following code is used to calculate the variance and standard deviation of the ‘Age’ variable for each category listed in the ‘Gender’ column:

#Print variance and sd of Age for each Gender

print(med_df.groupby('Gender')['Age'].agg([np.var, np.std]))

(2021)

The output is shown below:

	var	std
Gender		
Female	424.189726	20.595867
Male	427.130031	20.667124
Nonbinary	439.637423	20.967533

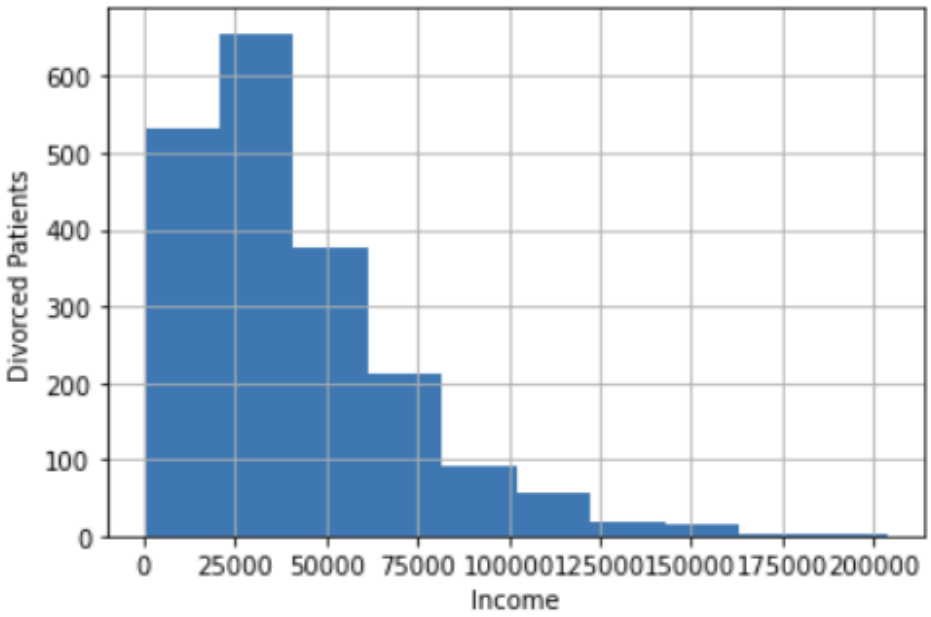
D1: Visual of findings

The following code is used to create a histogram which represents frequency distribution of income for marital status, ‘divorced’:

```
#Create a histogram of Income for Martial status 'Divorced'

med_df[med_df['Marital'] == 'Divorced']['Income'].hist()

plt.xlabel('Income')

plt.ylabel('Divorced Patients')

plt.show()
```

(DataCamp, 2021)

The output is shown below:

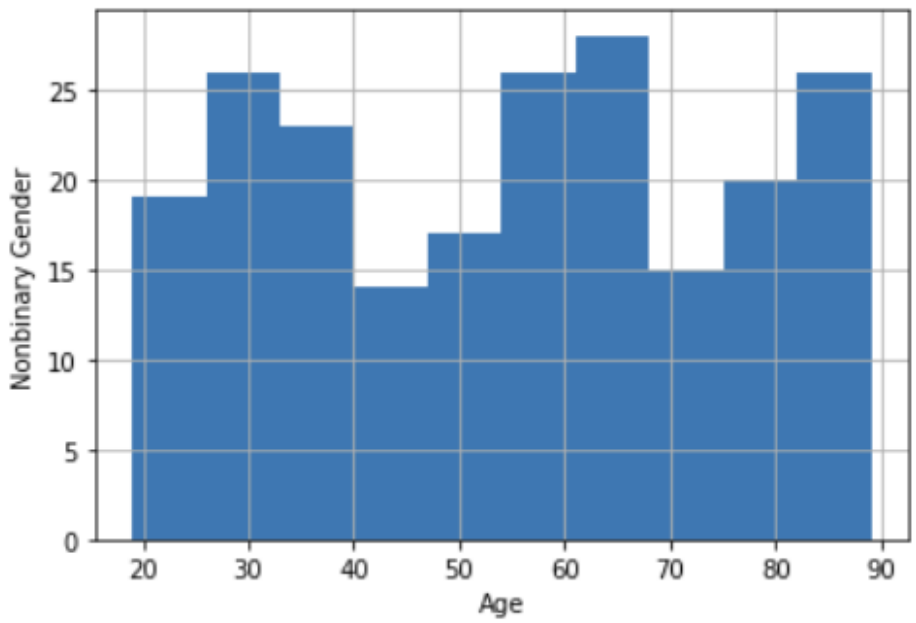


The following code is used to create a histogram which represents frequency distribution of age for gender specified as ‘nonbinary’:

```
#Create a histogram of Age for Gender specified as 'Nonbinary'

med_df[med_df['Gender'] == 'Nonbinary']['Age'].hist()

plt.xlabel('Age')

plt.ylabel('Nonbinary Gender')

plt.show()
```

(DataCamp, 2021)

The output is shown below:



E1: Results of analysis

```
The p-value approach to hypothesis testing in the decision rule
chisquare-score is: 0.7133125620168337  and p value is: 0.6128241714683622
Failed to reject the null hypothesis.
```

The output results from our chi-squared hypothesis testing are stated above. Since the p-value is ~0.6, which is lower than our chi-square-score of ~0.7, we cannot reject the null hypothesis. Recall from the beginning of this analysis, our alternative and null hypotheses state the following:

H1: The number of hospital readmissions differ between rural, suburban, and urban residential area types.

H0: The number of hospital readmissions do not differ between rural, suburban, and urban residential area types.

Since we fail to reject the null hypothesis, our original assumption that the number of hospital readmissions differ between rural, suburban, and urban residential area types is statistically insignificant with the sample data we have available. Even though the theory was proven to be incorrect, this is still critical information moving forward to help in ruling out potential causes of hospital readmission.

E2: Limitations of analysis

The main limitations of this analysis include a limited number of columns in the data set, in addition to limited ability to assess the scope of the data story over time. If the data set included columns with information to help better assess differences between patience such as quality of care, it would be easier to see clear differences between variables. Likewise, if the data set was accompanied by some time series data, we could further explore trends that may have been affected over time – painting a clearer picture of the story.

The chi-square test of independence also includes limitations of its own. While it can provide us with a p-value and analysis using a high degree of statistical accuracy, this specificity includes something to keep in mind. Due to its scrutinizing nature, the test tends to produce somewhat low correlation measures, even if the results are highly significant – and vice versa.

E3: Recommended course of action

Based on the overall analysis, I can conclude with 95% confidence that hospital readmission rates are not statistically influenced by residential area type at the population level. After conducting additional univariate and bivariate analyses on various attributes including: age, marital status, income, and gender, the sample data set does not reveal any significant questions of interest between these categories and their effect on hospital readmission rates. In response to this analysis, there are some actions that could be taken to study more possible correlations involving hospital readmission rates. For example, if we assume that hospital readmission rates might have something to do with the quality of hospital care, we can better analyze this possibility if additional data is provided to measure this variable of interest.

Web Sources

University of Portland Clark Library. (2021, October 8). APA Style (7th Edition) Citation Guide: Websites. LibGuides. https://libguides.up.edu/apa

The Python Software Foundation. (2021). Applications for Python. Python. Applications for Python | Python.org

Lumen Learning. (2013, September 17). Frequency Distributions for Quantitative Data. Boundless statistics. Frequency Distributions for Quantitative Data | Boundless Statistics (lumenlearning.com)

Statistical Language - Measures of Spread. (n.d.). Australian Bureau of Statistics. Retrieved December 27, 2021, from https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+measures+of+spread

Z. (2020, July 13). *How to Create Frequency Tables in Python*. Statology. Retrieved December 27, 2021, from https://www.statology.org/frequency-tables-python/

Data Camp. (n.d.). DataCamp. Retrieved November 20, 2021, from https://www.datacamp.com/users/sign_in?redirect=http%3A%2F%2Fapp.datacamp.com%2Flearn%2Fcustom-tracks%2Fcustom-d207-exploratory-data-analysis

Nigam, V. (2021, January 6). *Statistical Tests — When to use Which ? - Towards Data Science*. Medium. Retrieved December 22, 2021, from https://towardsdatascience.com/statistical-tests-when-to-use-which-704557554740