Performance Assessment: Exploratory Data Analysis

Joshua Funderburk
Western Governors University
D207: Exploratory Data Analysis
05/26/2021

A. Describe a real-world organizational situation or issue in the Data Dictionary:

1. **Question:** Which categorical variables are related to patient hospital readmission?
2. **Benefit from analysis:** An analysis of categorical variables and their relationship to hospital readmissions allows the organization to identify factors that lead to readmission. If a relationship is identified, the organization would be able to put policies and procedures in place during the initial stay to reduce the risk for readmission. Simply put, an analysis of this data could lead to the organization reducing readmissions and penalties that may occur because of readmissions.
3. **Data relevant to the question:**

Column Name	**Example**	**Data Type**
Gender	Male	Character
ReAdmis	No	Character
Soft_drink	No	Character
Initial_admin	Emergency Admission	Character
HighBlood	Yes	Character
Stroke	No	Character
Complication_risk	Medium	Character
Overweight	Yes	Character
Arthritis	Yes	Character
Diabetes	Yes	Character
Hyperlipidemia	No	Character
BackPain	Yes	Character
Anxiety	Yes	Character
Allergic_rhinitis	Yes	Character
Reflux_esophagitis	No	Character
Asthma	Yes	Character
Services	Blood Work	Character

B. Describe the data analysis by doing the following:

1. **Write code to run analysis of the data set:** See section B of the accompanying Jupyter Notebook for complete code. Example snip of Chi Square:

B1.1 Use Chi Square to test for a relationship between readmissions and gender

```
#Contingency table
contingency_table = pd.crosstab(data['ReAdmis'], data['Gender'])
print('Contingency Table = \n', contingency_table)

#Store contingency table values
observed_values = contingency_table.values

#Identify the test statistic, p-value, degrees of freedom, and expected values
stat, p, dof, expected = chi2_contingency(observed_values)

print('\nDegrees of Freedom =', dof)
print ('\nExpected Values =\n', expected)

#Interpret test statistic
prob = 0.95
critical = chi2.ppf(prob, dof)
print('Interpret Test Statistic:')
print('\nProbability = ', prob)
print('Critical Value = %.3f' % critical)
print('Test Statistic = %.3f' % stat)

if abs(stat) >= critical:
    print('\nOutcome: Dependent (reject H0)')
else:
    print('\nOutcome: Independent (fail to reject H0)')

#Interpret p-value
alpha = 1.0 - prob
print('Interpret P-Value:')
print('\nSignificance = %.3f' % alpha)
print('P-Value = %.3f' % p)

if p <= alpha:
    print('\nOutcome: Dependent (reject H0)')
else:
    print('\nOutcome: Independent (fail to reject H0)')
```

Source: Sewell, William

Source: Naik, Krish

Output and results of calculations: See section B of the accompanying Jupyter Notebook for complete code. Example snip of Chi Square results:

```
Contingency Table =
 Gender   Female  Male  Nonbinary
ReAdmis
No         3205  2995        131
Yes        1813  1773         83

Degrees of Freedom = 2

Expected Values =
 [[3176.8958 3018.6208  135.4834]
 [1841.1042 1749.3792   78.5166]]
Interpret Test Statistic:

Probability =  0.95
Critical Value = 5.991
Test Statistic = 1.586

Outcome: Independent (fail to reject H0)
Interpret P-Value:

Significance = 0.050
P-Value = 0.453

Outcome: Independent (fail to reject H0)
```

2. **Justification of analysis technique:** I chose to use Chi Square in my analysis because it allowed me to build one block of code and repeat it to test multiple categorical values against readmissions. Focusing in on categorical values with Chi Square gave me insights into the probability of relationship between Readmissions and sixteen other variables. Although ANOVA or a T-Test give great insight into non-categorical values, those statistical methods must be catered to each field and would not have allowed me to do such a broad analysis across the data set. However, Performing ANOVA and T-Tests for the remaining non-categorical fields would be wise to ensure all relationships to hospital Readmissions are uncovered. For the purpose of this analysis and assessment requirements, Chi Square was the best place to start to establish a foundation for relationships. Additionally, Chi Square was the best technique to answer the question posed in Section A1 due to the technique's purpose of analyzing categorical values.

C. Identify the distribution of two continuous variables and two categorical variables using univariate statistics:

1. **Represent the findings in Part C:** See section C of the accompanying Jupyter Notebook for complete code. Code for section C1:
    i. **Age Statistics – Continuous Variable 1**

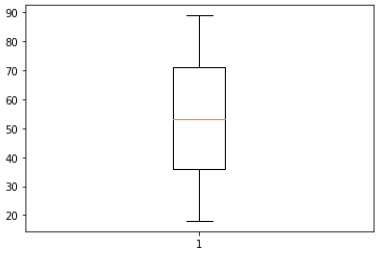
Age Statistics (Continuous Variable)

```
In [21]: data['Age'].describe()

Out[21]: count    10000.000000
         mean        53.511700
         std         20.638538
         min         18.000000
         25%         36.000000
         50%         53.000000
         75%         71.000000
         max         89.000000
         Name: Age, dtype: float64

In [22]: plt.boxplot(data.Age)

Out[22]: {'whiskers': [<matplotlib.lines.Line2D at 0x1e062aad9c8>,
           <matplotlib.lines.Line2D at 0x1e062aadf88>],
          'caps': [<matplotlib.lines.Line2D at 0x1e062aadc88>,
           <matplotlib.lines.Line2D at 0x1e062ab8a08>],
          'boxes': [<matplotlib.lines.Line2D at 0x1e062aad2c8>],
          'medians': [<matplotlib.lines.Line2D at 0x1e062ab8fc8>],
          'fliers': [<matplotlib.lines.Line2D at 0x1e062aa39c8>],
          'means': []}
```



ii. Doctor Visit Statistics – Continuous Variable 2

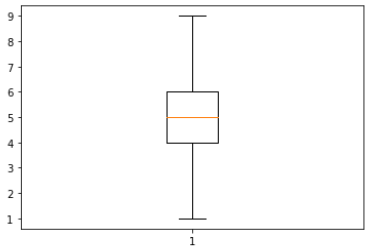
Doctor Visit Statistics (Continuous Variable)

```
In [23]: data['Doc_visits'].describe()

Out[23]: count    10000.000000
         mean         5.012200
         std          1.045734
         min          1.000000
         25%          4.000000
         50%          5.000000
         75%          6.000000
         max          9.000000
         Name: Doc_visits, dtype: float64

In [24]: plt.boxplot(data.Doc_visits)

Out[24]: {'whiskers': [<matplotlib.lines.Line2D at 0x1e062b60cc8>,
           <matplotlib.lines.Line2D at 0x1e062b60d88>],
          'caps': [<matplotlib.lines.Line2D at 0x1e062b66a88>,
           <matplotlib.lines.Line2D at 0x1e062b66f48>],
          'boxes': [<matplotlib.lines.Line2D at 0x1e062b60b08>],
          'medians': [<matplotlib.lines.Line2D at 0x1e062b66c08>],
          'fliers': [<matplotlib.lines.Line2D at 0x1e062b6c948>],
          'means': []}
```



iii. Services Statistics - Categorical Variable 1

Source: Piush

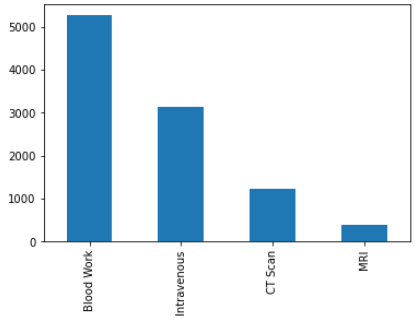
Services Statistics (Categorical Variable)

```
In [25]: data['Services'].describe()

Out[25]: count          10000
         unique             4
         top       Blood Work
         freq            5265
         Name: Services, dtype: object

In [26]: data['Services'].value_counts().plot.bar()

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x1e062bed908>
```

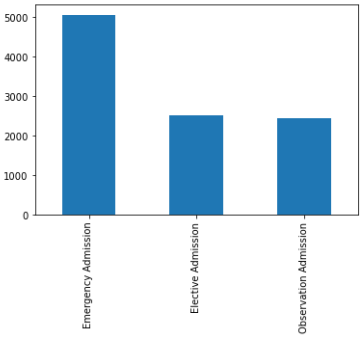


iv. **Initial Admission Reason Statistics – Categorical Variable 2**
Source: Piush

Initial Admission Reason Statistics (Categorical Variable)

```
In [27]: data['Initial_admin'].describe()
Out[27]: count                   10000
         unique                      3
         top       Emergency Admission
         freq                     5060
         Name: Initial_admin, dtype: object

In [28]: data['Initial_admin'].value_counts().plot.bar()
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x1e062c8c788>
```



D. Identify the distribution of two continuous variables and two categorical variables using bivariate statistics:

1. **Represent the findings in Part D:** See section D of the accompanying Jupyter Notebook for complete code.

i. **Age vs Readmissions – Continuous Variable 1**

Age vs ReAdmissions (Continuous Variable)

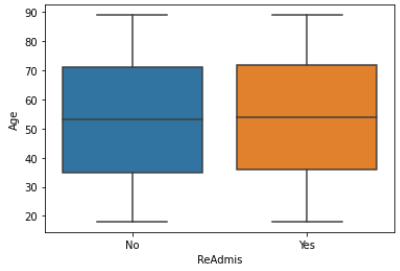
```
In [29]: ReAdmis_Age_Grouped = pd.crosstab(index=data['ReAdmis'], columns=data['Age'])
         ReAdmis_Age_Grouped
```

Out[29]:

Age	18	19	20	21	22	23	24	25	26	27	...	80	81	82	83	84	85	86	87	88	89
ReAdmis																					
No	85	92	86	81	88	98	84	87	95	79	...	71	85	74	86	89	72	91	91	83	86
Yes	48	45	34	44	53	39	60	43	49	56	...	45	46	50	48	38	63	65	45	60	46

2 rows × 72 columns

```
In [30]: sns.boxplot(x='ReAdmis', y='Age', data=data)
         plt.show()
```



ii. # of Doctor Visits vs Readmissions – Continuous Variable 2

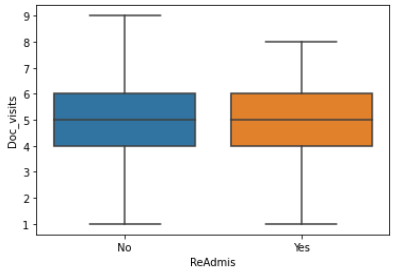
# of Doctor Visits vs ReAdmissions (Continuous Variable)

In [31]:
```
ReAdmis_DoctorVisits_Grouped = pd.crosstab(index=data['ReAdmis'], columns=data['Doc_visits'])
ReAdmis_DoctorVisits_Grouped
```

Out[31]:

Doc_visits	1	2	3	4	5	6	7	8	9
ReAdmis									
No	4	36	375	1511	2416	1558	392	37	2
Yes	2	22	220	874	1407	878	242	24	0

In [32]:
```
sns.boxplot(x='ReAdmis', y='Doc_visits', data=data)
plt.show()
```



iii. Services vs Readmissions – Categorical Variable 1

Services vs ReAdmissions (Categorical Variable)

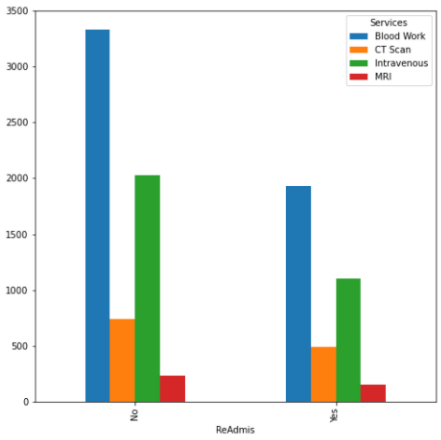
In [35]:
```
ReAdmis_Services_Grouped = pd.crosstab(index=data['ReAdmis'], columns=data['Services'])
ReAdmis_Services_Grouped
```

Out[35]:

Services	Blood Work	CT Scan	Intravenous	MRI
ReAdmis				
No	3335	737	2027	232
Yes	1930	488	1103	148

In [36]:
```
ReAdmis_Services_Grouped.plot(kind="bar", figsize=(8,8))
```

Out[36]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e0634819c8>
```



iv. **Complication Risk vs Readmissions – Categorical Variable 2**

Complication Risk vs ReAdmissions (Categorical Variable)

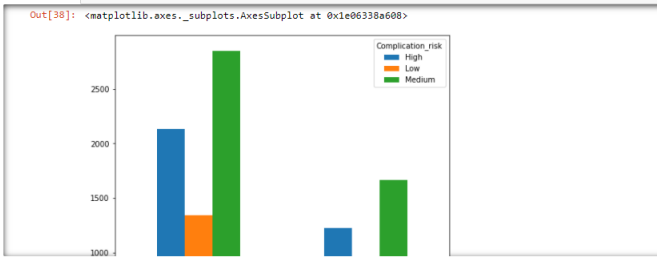
```
In [37]: ReAdmis_CompRisk_Grouped = pd.crosstab(index=data['ReAdmis'], columns=data['Complication_risk'])
         ReAdmis_CompRisk_Grouped
```

Out[37]:

ReAdmis \ Complication_risk	High	Low	Medium
No	2135	1343	2853
Yes	1223	782	1664

```
In [38]: ReAdmis_CompRisk_Grouped.plot(kind="bar", figsize=(8,8))
```

```
Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0x1e06338a608>
```



E. Summary of the data analysis implications:

1. **Results of the hypothesis test:** Of the sixteen categorical variables analyzed for a relationship to hospital readmissions, the null hypothesis was only rejected for one variable. The Chi Square tests revealed that there is a dependency between readmissions and the services rendered during the initial hospital stay. The following variables were determined to be independent of readmissions: gender, soft drinks, initial admissions reason, high blood pressure, stroke, complication risk level, overweight, arthritis, diabetes, hyperlipidemia, back pain, anxiety, allergic rhinitis, reflux esophagitis, and asthma.
2. **Limitations of the data analysis:**
    i. A major limitation of this data analysis, and hypothesis testing in general, is that the test does not explain the reason as to why a difference exists ("Limitations of Hypothesis testing in Research"). The results of this analysis simply identify where there are differences – further analysis and consultation with subject matter experts is required to understand why there are differences.
    ii. The results of this analysis are based on probabilities ("Limitations of Hypothesis testing in Research"). There cannot be absolute certainty in the results.
3. **Recommended course of action:**
    i. It is recommended that the relationship between services during the initial stay and readmissions be explored in more detail. A great start would be to look for relationships between the types of services (MRI, Blood Work, etc.) and readmissions to try to identify any type of service that may indicate risk for readmissions. Subject matter experts should also be enlisted to further examine the relationship.
    ii. It is imperative to find relationships to readmissions. Thus, re-running the Chi Square tests with a higher alpha, although potentially less accurate, may help find further relationships between categorical values.

iii. Statistical analysis should be expanded beyond categorical values to identify relationships between available data and readmissions.

Works Cited

Bruce, P.A. (2020). *Practical Statistics for Data Scientists, 50 Essential Concepts Using R and Python.* Sebastopol, CA: O'Reilly Media, Incorporated. ISBN: 978-1792072942

Data to Fish. (n.d.). Retrieved from https://datatofish.com/round-values-pandas-dataframe/

Limitations of Hypothesis testing in Research. (n.d.). Retrieved May 20, 2021, from https://www.wisdomjobs.com/e-university/research-methodology-tutorial-355/limitations-of-the-tests-of-hypotheses-11539.html

Naik, Krish. (2020). *Tutorial 33- Chi Square Test Implementation with Python- Hypothesis Testing-Part 2* [Video]. Retrieved 20 May 2021, from https://www.youtube.com/watch?v=w5iKu1IrTJQ.

Pandas.crosstab. (n.d.). Retrieved from https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html

Piush, Vaish. (2021, May 15). *Visualise Categorical Variables in Python.* Retrieved from https://adataanalyst.com/data-analysis-resources/visualise-categorical-variables-in-python/

Sewell, William (n.d.). *Chi-Square for EDA* D207 [Video]. Retrieved May 22, 2021, from https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=52d9e72f-3309-4780-ac2b-accf014a436f