D207 – Exploratory Data Analysis

Task 1

William Stults

03/12/2022

Table of Contents

Contents

William Stults – D207

Description of Situation and Data Set

Research Question

My dataset for this data cleaning exercise includes data on an internet service provider's current and former subscribers, with an emphasis on customer churn (whether customers are maintaining or discontinuing their subscription to the ISP's service). Data analysis performed on the dataset will be aimed with this research question in mind: is there a relationship between customer lifestyle, or "social" factors, and customer churn? Lifestyle and social factors might include variables such as age, income, and marital status, among others.

Benefit to Stakeholders

Conclusions gleaned from analysis of this data can benefit stakeholders by revealing information on which customer populations may be more likely to "churn", or to terminate their service contract with the ISP. Such information may be used to fuel targeted advertising campaigns, special promotional offers, and other strategies related to customer retention.

Description of Variables

The table below describes the variables included in the customer churn dataset that are relevant to the research question.

Variable	**Data Type**	**Quantitative / Qualitative**	**Description**
Area	object	Qualitative	Type of area customer lives in (urban, suburban, rural)
Children	float64	Quantitative	How many children live in the customer's household
Age	float64	Quantitative	Customer's age
Income	float64	Quantitative	Customer's income annually
Marital	object	Qualitative	Customer's marital status
Gender	object	Qualitative	Self-identified gender of the customer
Churn	object	Qualitative	Yes/No if customer canceled service

Description of Data Analysis

Analysis of Data Using Chi-square

I will use the Chi-square test of independence for my analysis of the categorical variables in my data set. To do so, I will need to generate contingency tables that include the counts of each variable when evaluated against my dependent variable's values. The "Churn" variable will always contain one of 2 values: "Yes" or "No". I will use the crosstab() function from the Pandas library in my creation of the tables. The code used to create my contingency tables is shown below.

```
## Generate contingency tables for each categorical variable

ct_Marital = pd.crosstab(data_slice.Churn, data_slice.Marital,
margins=True)

ct_Gender = pd.crosstab(data_slice.Churn, data_slice.Gender,
margins=True)

ct_Area = pd.crosstab(data_slice.Churn, data_slice.Area, margins=True)
```

My null hypothesis will be that, using a significance level of 0.05, there is no statistically significant relationship between each set of variables evaluated. With contingency tables created, my next step will be to leverage the array() function from NumPy in conjunction with the chi2_contingency() function from the Stats module of the SciPy library to calculate the Chi-square statistic, the p-value, and the degree of freedom for each combination of variables. The code used for this operation is shown below.

```
## Generate Chi-square, p-value and degree of freedom for each
variable pair

obs = np.array([ct_Marital.iloc[0][0:5].values,
ct_Marital.iloc[1][0:5].values])

stats.chi2_contingency(obs)[0:3]

obs = np.array([ct_Gender.iloc[0][0:3].values,
ct_Gender.iloc[1][0:3].values])

stats.chi2_contingency(obs)[0:3]

obs = np.array([ct_Area.iloc[0][0:3].values,
ct_Area.iloc[1][0:3].values])

stats.chi2_contingency(obs)[0:3]
```

William Stults – D207

Results and Conclusions

Pictured below are the resulting contingency tables for each variable.

Marital –

Marital Churn	Divorced	Married	Never Married	Separated	Widowed	All
No	1539	1418	1468	1454	1471	7350
Yes	553	493	488	560	556	2650
All	2092	1911	1956	2014	2027	10000

Gender –

Gender Churn	Female	Male	Nonbinary	All
No	3753	3425	172	7350
Yes	1272	1319	59	2650
All	5025	4744	231	10000

Area –

Area Churn	Rural	Suburban	Urban	All
No	2464	2473	2413	7350
Yes	863	873	914	2650
All	3327	3346	3327	10000

William Stults – D207

The images below reveal the results of the Chi-square, p-value, and degree of freedom for each variable.

Marital –

```
obs = np.array([ct_Marital.iloc[0][0:5].values, ct_Marital.iloc[1][0:5].values])
stats.chi2_contingency(obs)[0:3]
```

```
(5.565780556713389, 0.23400754115227573, 4)
```

Gender –

```
obs = np.array([ct_Gender.iloc[0][0:3].values, ct_Gender.iloc[1][0:3].values])
stats.chi2_contingency(obs)[0:3]
```

```
(7.880065153719115, 0.019447581193944605, 2)
```

Area –

```
obs = np.array([ct_Area.iloc[0][0:3].values, ct_Area.iloc[1][0:3].values])
stats.chi2_contingency(obs)[0:3]
```

```
(2.4390738588073266, 0.2953669109921032, 2)
```

Below are the resulting values, organized as a table:

Variable	**Chi-square value**	**p-value**	**degree of freedom**
Marital	5.565780556713389	0.23400754115227573	4
Gender	7.880065153719115	0.019447581193944605	2
Area	2.4390738588073266	0.2953669109921032	2

Considering the significance level of .05, the only variable in the set which cannot reject the null hypothesis is Gender, with a p-value of 0.019447581193944605. All p-values for the other variables in the set are greater than the significance level, and reject the null hypothesis, indicating we cannot conclusively say there is no relationship between each of those variables and Churn.

Justification of Approach

All code execution was carried out via Jupyter Lab, using Python 3. I used Python as my selected programming language due to prior familiarity and broader applications when considering programming in general. R is a very strong and robust language tool for data analysis and statistics but finds itself somewhat limited to that niche role (Insights for Professionals, 2019). I utilized the NumPy, Pandas, and Matplotlib libraries to perform many of my data analysis tasks, as they are among the most popular Python libraries employed for this purpose and see widespread use. Seaborn is included primarily for its better-looking boxplots, seen later in this document (Parra, 2021). Beyond these libraries, SciPy builds upon NumPy, incorporating NumPy's many mathematical functions and including numerous others (Johari, 2019). It was used here primarily for its stats module.

The Chi-square Test of Independence is intended for use when evaluating two categorical or nominal variables to prove or disprove whether a relationship between the two exists. Given the dependent variable, Churn, is categorical, it was the best choice for this exercise. A significance level of 0.05 is recognized to generally work well when determining whether variables are independent. This significance level "indicates a 5% risk of concluding that an association between the variables exists when there is no actual association" (Minitab, 2022).

While the T-Test and ANOVA test are also powerful tools for these types of comparisons, they are intended for use with numerical data primarily. As the dependent variable is not numeric, using either of these tests would not have been an effective choice.

Univariate Statistics

Identifying Distributions

I observed the distributions of two continuous variables, Age and Income, via the hist() function from the MatPlotLib library's PyPlot module. The code used is shown below.

```
## Univariate analysis of Age and Income via histogram
plt.hist(data_slice['Age'])
plt.hist(data_slice['Income'])
```

The resulting histograms revealed that the Age variable has a normal distribution with ages ranging between 18 and 89. The histogram for income has a left-skewed distribution, also known as an F distribution, with a range starting at 348.67 and extending up to 258900.7.

For univariate analysis of categorical variables Gender and Marital, I used Seaborn's countplot() function to create vertical bar charts. The code for these operations is shown below.

```
## Univariate analysis of Gender and Marital via barchart
sns.countplot(x="Gender", data=data_slice)
sns.countplot(x=" Marital", data=data_slice)
```

The bar chart for Gender reveals a nearly even amount of Male and Female, with very few Nonbinary. For Marital, the bar chart indicates a normal distribution among all marital statuses.

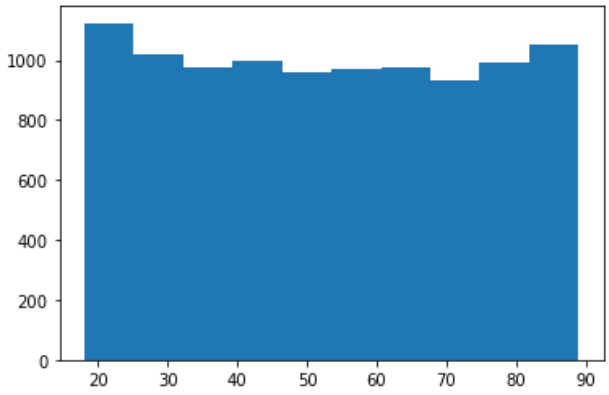
William Stults – D207

Visual Representation of Findings

Histogram – Age

```
plt.hist(data_slice['Age'])
```

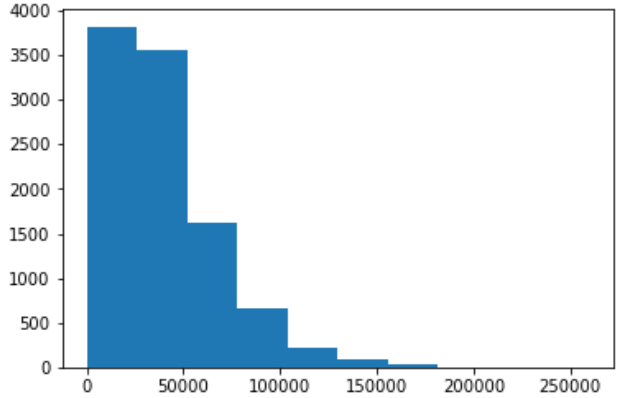
```
(array([1124., 1019.,  976.,  997.,  961.,  971.,  974.,  935.,  992.,
        1051.]),
 array([18. , 25.1, 32.2, 39.3, 46.4, 53.5, 60.6, 67.7, 74.8, 81.9, 89. ]),
 <BarContainer object of 10 artists>)
```



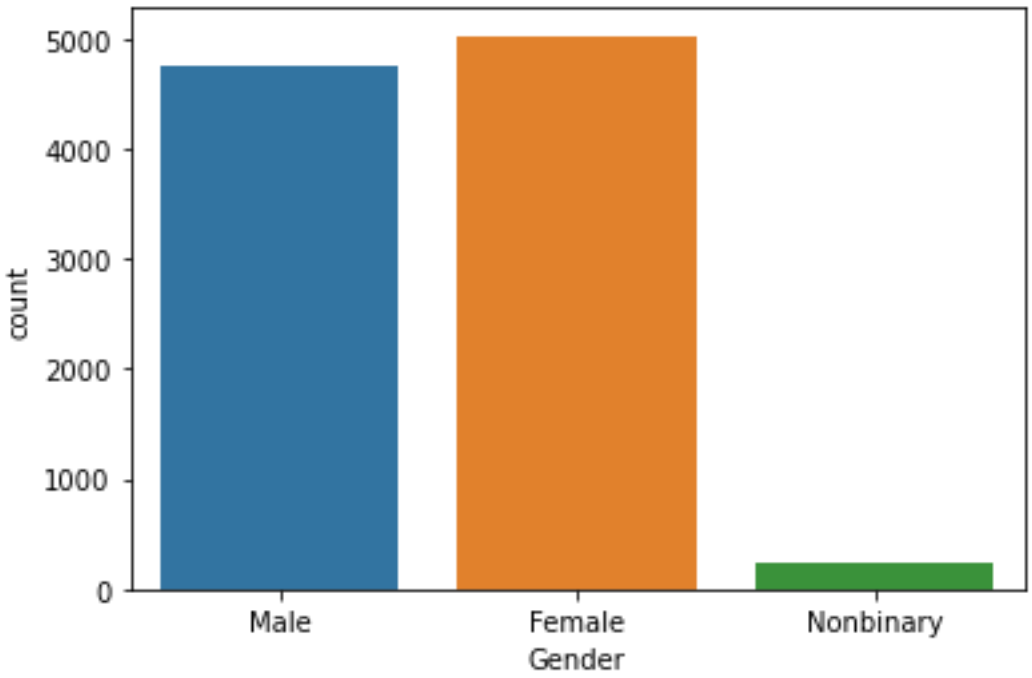
Histogram – Income

```
plt.hist(data_slice['Income'])
```

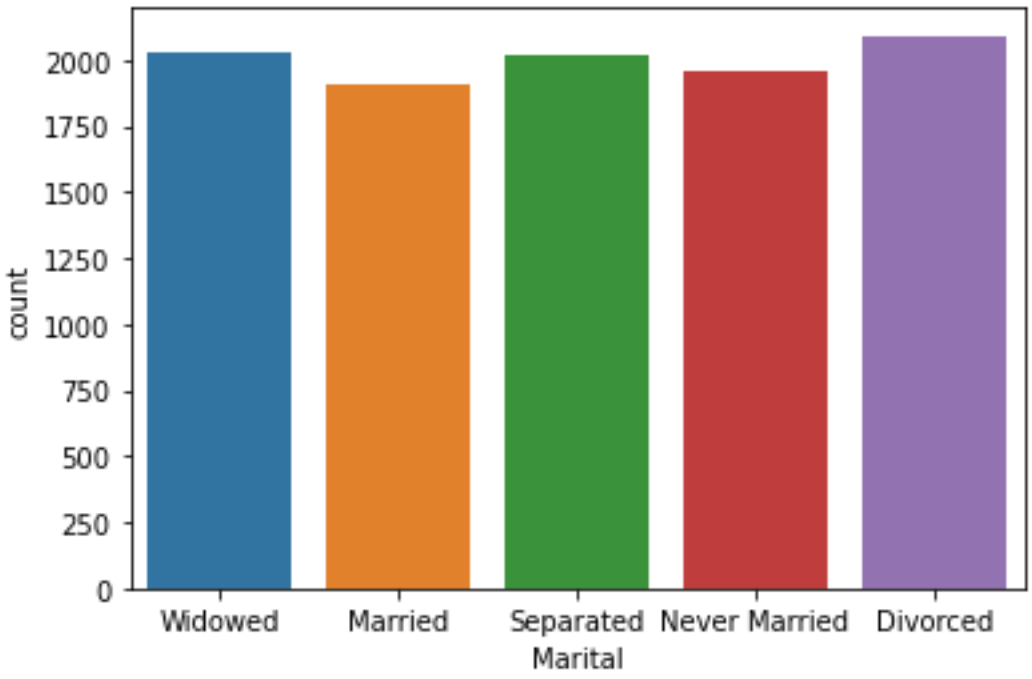
```
(array([3.822e+03, 3.560e+03, 1.624e+03, 6.570e+02, 2.200e+02, 8.100e+01,
        2.600e+01, 5.000e+00, 3.000e+00, 2.000e+00]),
 array([   348.67 ,  26203.873,  52059.076,  77914.279, 103769.482,
        129624.685, 155479.888, 181335.091, 207190.294, 233045.497,
        258900.7  ]),
 <BarContainer object of 10 artists>)
```



Bar chart - Gender



Bar chart - Marital



Bivariate Statistics

Identifying Distributions

I used the scatterplot() function from Seaborn to determine if any correlation exists between the continuous variables Age and Income. Here is the code that was used:

```
## Scatterplot revealing correlation between Age and Income
biv_cont_data = data_slice[['Age', 'Income']]
sns.scatterplot(x="Age", y="Income", data=biv_cont_data)
```

The scatterplot revealed no apparent correlation between Age and Income. Incomes of varying amounts are distributed somewhat randomly above a value near 100,000 but do not appear to favor a particular Age value.

Distribution of Gender and Marital together was evaluated via a stacked bar chart, using Seaborn's displot() function. The code is shown below.

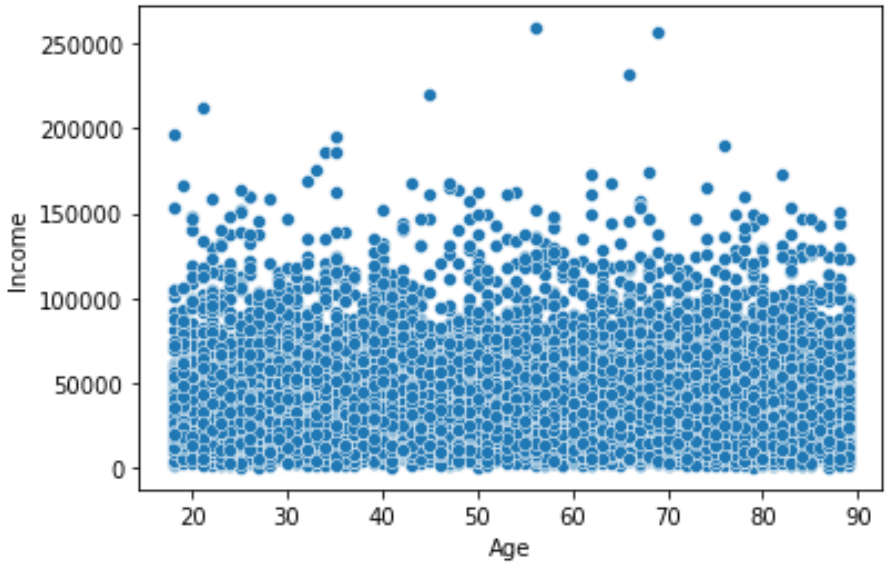
```
## Stacked bar chart revealing correlation between Gender and Marital
biv_cat_data = data_slice[['Gender', 'Marital']]
sns.displot(biv_cat_data, x='Marital', hue='Gender', multiple='stack')
```

Again, no correlation is apparent from the result, with each Marital value displaying a near equal proportional distribution between Male and Female, with Nonbinary in far fewer numbers.

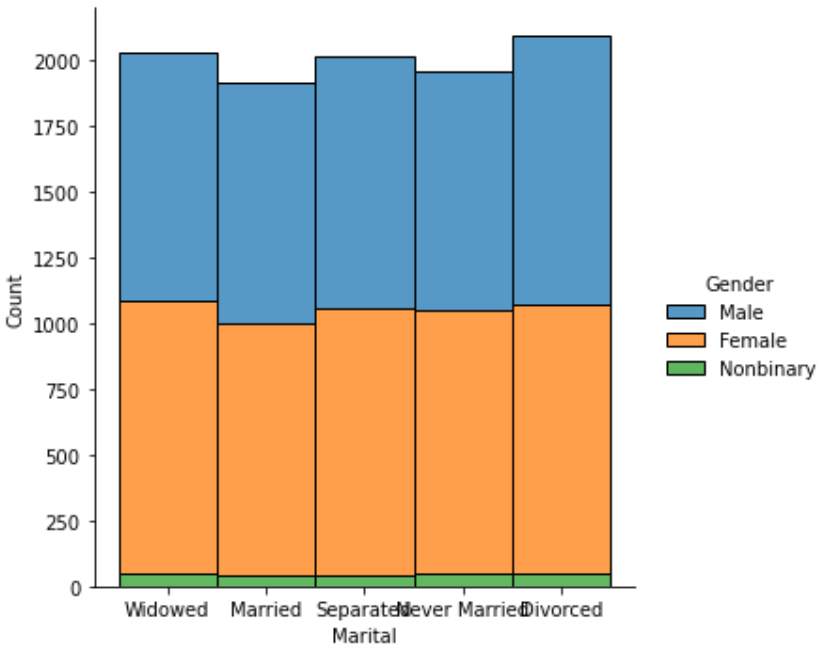
William Stults – D207

Visual Representation of Findings

Scatterplot – Income and Age



Stacked bar chart – Gender and Marital



Overall Results

Results of Analysis

Based upon the analysis outlined above, the only variable amongst the group of categorical variables evaluated which confirmed the null hypothesis was Gender. All other variables rejected the null hypothesis. We can conclude that there may be statistically significant relationships between those social characteristics and whether a customer will choose to terminate their service contract with the ISP.

I also determined the distributions of Age and Income among the population sample, with customers' ages spread rather homogenously between ages 18 and 89, while Income figures largely favor subjects earning less than $100,000 per year. The population sample represents a fairly even split between subjects who responded to the Gender prompt, showing us an almost equal amount of Males and Females, while Marital statuses are represented equally.

Limitations of Analysis

I have three variables that were not categorical. Having used only one of the testing methods available, only a partial evaluation of my complete set of variables was possible. While the results of the testing were enough to indicate that relationships likely exist between customer social factors and customer retention, whether those relationships are positive or negative would require further analysis.

Recommendations

I began my data exploration exercise aiming to determine whether there is a relationship between customer lifestyle, or "social" factors, and customer churn. Based upon the insight gained from performing the analysis of this data, the answer to that question is clearly yes. The next recommended actions to be taken would be to perform further analysis to conclude which of these social factors have the greatest and least impact on customer retention, both positively and negatively. Having those conclusions available would empower the business with actionable insights to better

retain customer business, and potentially target advertising and promotions at specific populations in an attempt to better strengthen the ISP's relationship with its customers.

Web Source References

https://towardsdatascience.com/chi-square-test-for-independence-in-python-with-examples-from-the-ibm-hr-analytics-dataset-97b9ec9bb80a

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html

https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html

https://www.sfu.ca/~mjbrydon/tutorials/BAinPy/08_correlation.html

https://stackoverflow.com/questions/50319614/count-plot-with-stacked-bars-per-hue

https://pythonbasics.org/seaborn-barplot/

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.bar.html

References

Insights for Professionals. (2019, February 26). *5 Niche Programming Languages (And Why They're Underrated)*. **https://www.insightsforprofessionals.com/it/software/niche-programming-languages**

Johari, A. (2019, September 4). *What is Python SciPy and How to use it?*. Medium. **https://medium.com/edureka/scipy-tutorial-38723361ba4b**

Minitab. (2022). Interpret the key results for Chi-Square Test for Association. **https://support.minitab.com/en-us/minitab/19/help-and-how-to/statistics/tables/how-to/chi-square-test-for-association/interpret-the-results/key-results/**

Parra, H. (2021, April 20). *The Data Science Trilogy*. Towards Data Science. **https://towardsdatascience.com/the-data-science-trilogy-numpy-pandas-and-matplotlib-basics-42192b89e26**