Performance Assessment:

Exploratory Data Analysis

Ryan James Calabio

D207: Exploratory Data Analysis

Apr 29, 2024

A1: Question For Analysis

The question I intend to ask for this analysis is “whether tenure has a significant relationship with churn”? We can test this using ANOVA to test the relationship for significance of quantitative variables to churn. This means our null and alternative hypothesis are as follows:

H_0: there is no significant relationship between the variable and churn (the means of the categories of the quantitative variables are equal)

H_a: there is a significant relationship between the variable and churn

I will use a confidence level of 95%.

A2: Benefit From Analysis

Stakeholders can benefit from this analysis because understanding whether or not tenure will affect churn can help them know what kind of business strategy they should focus on. If tenure has a significant relationship with churn, then one can also attempt to understand what variables are related to tenure and churn, and then get a better understanding of how the variables interactions may affect the outcome.

A3: Data Identification

The variables that I use in this analysis are churn, tenure, income, MonthlyCharge, marital, and gender. Churn, income, and tenure are used in the ANOVA analysis. Churn, tenure, income, MonthlyCharge, marital, and gender are used in bivariate visualizations.

B1: Code

See code included.

B2: Output

ANOVA Analysis

```
Variable:  MonthlyCharge
 F_onewayResult(statistic=1615.1940392182648, pvalue=0.0)

Variable:  Income
 F_onewayResult(statistic=0.3524671270244696, pvalue=0.5527332919027459)

Variable:  Tenure
 F_onewayResult(statistic=3083.011240356265, pvalue=0.0)
```

B3: Justification

I chose the ANOVA because it allows us to test quantitative dependent variables influence on independent categorical variables. I tested Tenure, MonthlyCharge, and Income using an ANOVA for loop. My results concluded that MonthlyCharge and Tenure have a significant relationship with churn as their p-value is below .05 so we can reject the null hypothesis that the variables are not significantly related with churn. Income has a p-value of .55 so we have to accept the null hypothesis for that variable. This gives us insight into our research question as it answers it and gives substantial evidence that there is a significant relationship between churn and tenure.

C: Univariate Statistics

The two continuous variables I chose to create univariate visuals for are tenure and income. The two categorical variables I chose to create univariate visuals for are churn and gender. The univariate graph for tenure I created is a histogram. Looking at the histogram, we can see that the distribution of the variable is bimodal. Using the describe function, we can delve more into the univariate statistics of tenure. We can see that the average length of a customer staying with this business is 34.5 years. The standard deviation is 26.44 years. The minimum length of a recorded customer's tenure is 1 year and the longest is 71.99.

```
df['Tenure'].describe()
```

```
count    10000.000000
mean        34.526188
std         26.443063
min          1.000259
25%          7.917694
50%         35.430507
75%         61.479795
max         71.999280
Name: Tenure, dtype: float64
```

The univariate graph for income I created is a histogram. From the histogram, the distribution is skewed right. Looking at the describe function below, we can delve more into the univariate statistics related to income. We can see that customers have a mean income of 39,806 dollars and a standard deviation of 28,199 dollars. The minimum income of a customer is 348 dollars, and the max is 258,900 dollars.

```
df['Income'].describe()
```

```
count     10000.000000
mean      39806.926771
std       28199.916702
min         348.670000
25%       19224.717500
50%       33170.605000
75%       53246.170000
max      258900.700000
Name: Income, dtype: float64
```

The univariate graph for gender I created is a bar plot. This shows us the categorical distribution that most customers are women, then men, and then non-binary in order of most to least amount. We can use the describe function to delve into the univariate statistics for gender, which has 3 unique categories. The category with the highest count is “Female” which has 5025 customers.

```
df['Gender'].describe()
```

```
count      10000
unique         3
top       Female
freq        5025
Name: Gender, dtype: object
```

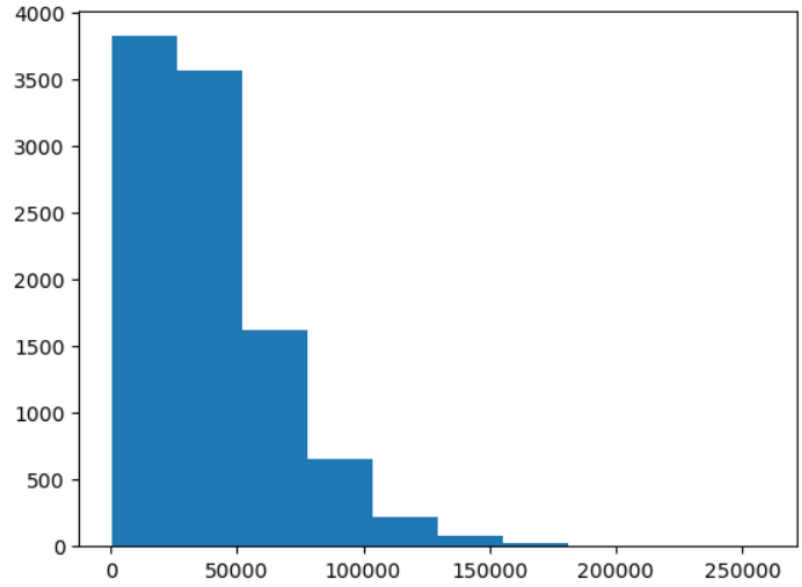
The univariate graph for churn I created is a bar plot. It shows us the data distribution of the churn variable, and that most customers do not churn and are retained. However, there is still a large number of customers that do churn. We can use the describe function again to go further into the univariate statistics of churn. We can see that there are two unique categories. The top category is “No” which includes 7,350 customers under this category.

```
df['Churn'].describe()
```

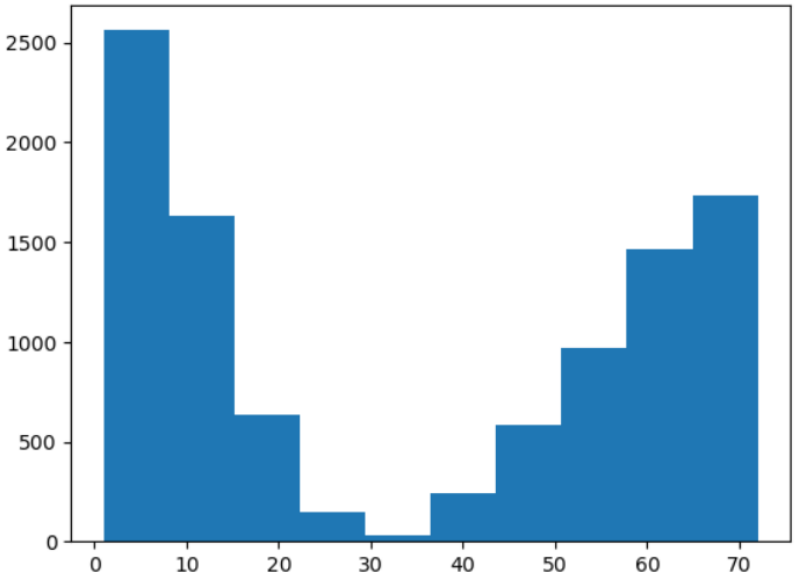
```
count     10000
unique        2
top          No
freq       7350
Name: Churn, dtype: object
```

C1: Visual of Findings

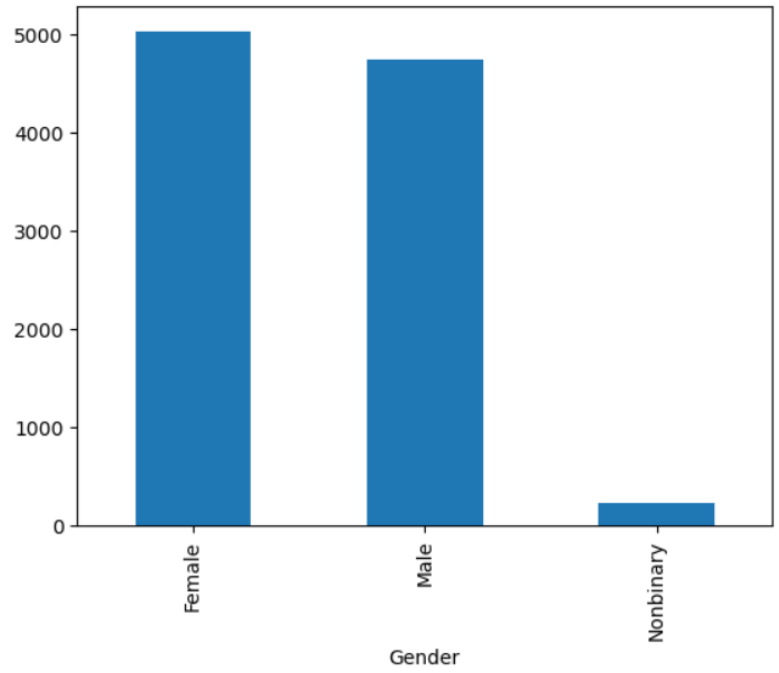
Income Histogram created using matplotlib:



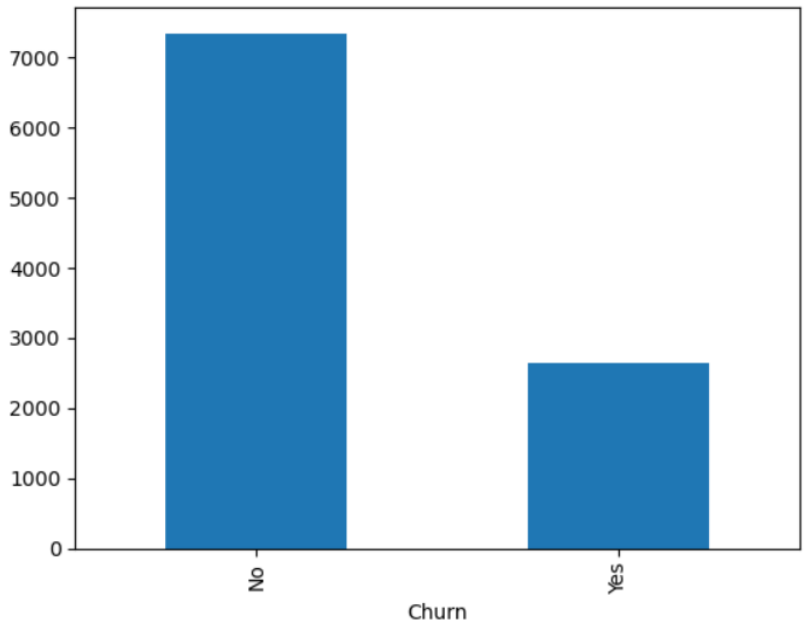
Tenure histogram created using matplotlib:



Gender Bar Plot



Churn Bar Plot



D: Bivariate Statistics

The first bivariate scatterplot graph I created uses two quantitative variables and that is Tenure and MonthlyCharge. This graph helps us understand if there's any kind of linear correlation between the two quantitative variable or if it's just noise. Looking at the graph it's hard to see a correlation between the two variables. There is also an empty space between 20 and 40. The mean for MonthlyCharge is 172.62, the standard deviation is 42.94, minimum is 79.97, and max is 290.16. We can also look at our bivariate statistics of our ANOVA anlysis using tenure and churn, and monthly churn and tenure. From our output of that analysis, we know that both of these variables have a significant relationship with churn.

```
count    10000.000000
mean       172.624816
std         42.943094
min         79.978860
25%        139.979239
50%        167.484700
75%        200.734725
max        290.160419
Name: MonthlyCharge, dtype: float64
```

The second bivariate graph that I created is a tenure and income scatterplot. In this scatterplot we can see that income tends to be below 200k with it popping above that in a few outliers. One can a large empty space between a tenure of 20 and 40 which could be something interesting to explore. This same empty space appeared in our graph with Tenure and MonthlyCharge, so it seems like there is a lot of missing data for tenure between values and 20 and 40. We can again look at the bivariate statistics of our ANOVA analysis to understand that tenure is significantly related to churn but income is not. This can give context to our scatterplot, as we can focus on how variables that are not related to churn may affect variables that are. We can also look deeper at the summary statistics again for tenure and income.

```
df['Income'].describe()
```

```
count     10000.000000
mean      39806.926771
std       28199.916702
min         348.670000
25%       19224.717500
50%       33170.605000
75%       53246.170000
max      258900.700000
Name: Income, dtype: float64
```

```
df['Tenure'].describe()
```

```
count    10000.000000
mean        34.526188
std         26.443063
min          1.000259
25%          7.917694
50%         35.430507
75%         61.479795
max         71.999280
Name: Tenure, dtype: float64
```

Comparing these two summary statistics can give us an understanding of the scale of the scatterplot and what to expect for our comparison. For instance, we can know that our x-axis will not go past 72 as that is the max for tenure and will not go below 1 as that is the minimum.

The bivariate graph for marital and churn I created is a stacked bar chart, and it points towards the conclusion that there is not much of an effect on the churn based on marriage status. This is because most of the different counts don't change a significant amount based on marriage status but instead on churn. We can go into marital further here using the describe function. There are five categories in marital, and the top category is "Divorced". This category has 2,092 customers included in it.

```
df['Marital'].describe()
```

```
count        10000
unique           5
top       Divorced
freq          2092
Name: Marital, dtype: object
```

On top of this, we can also look into the bivariate statistics of marital and churn by looking at the crosstab. Here we can see the exact counts of each category of marital by each category of churn.

Marital Churn	Divorced	Married	Never Married	Separated	Widowed
No	1539	1418	1468	1454	1471
Yes	553	493	488	560	556

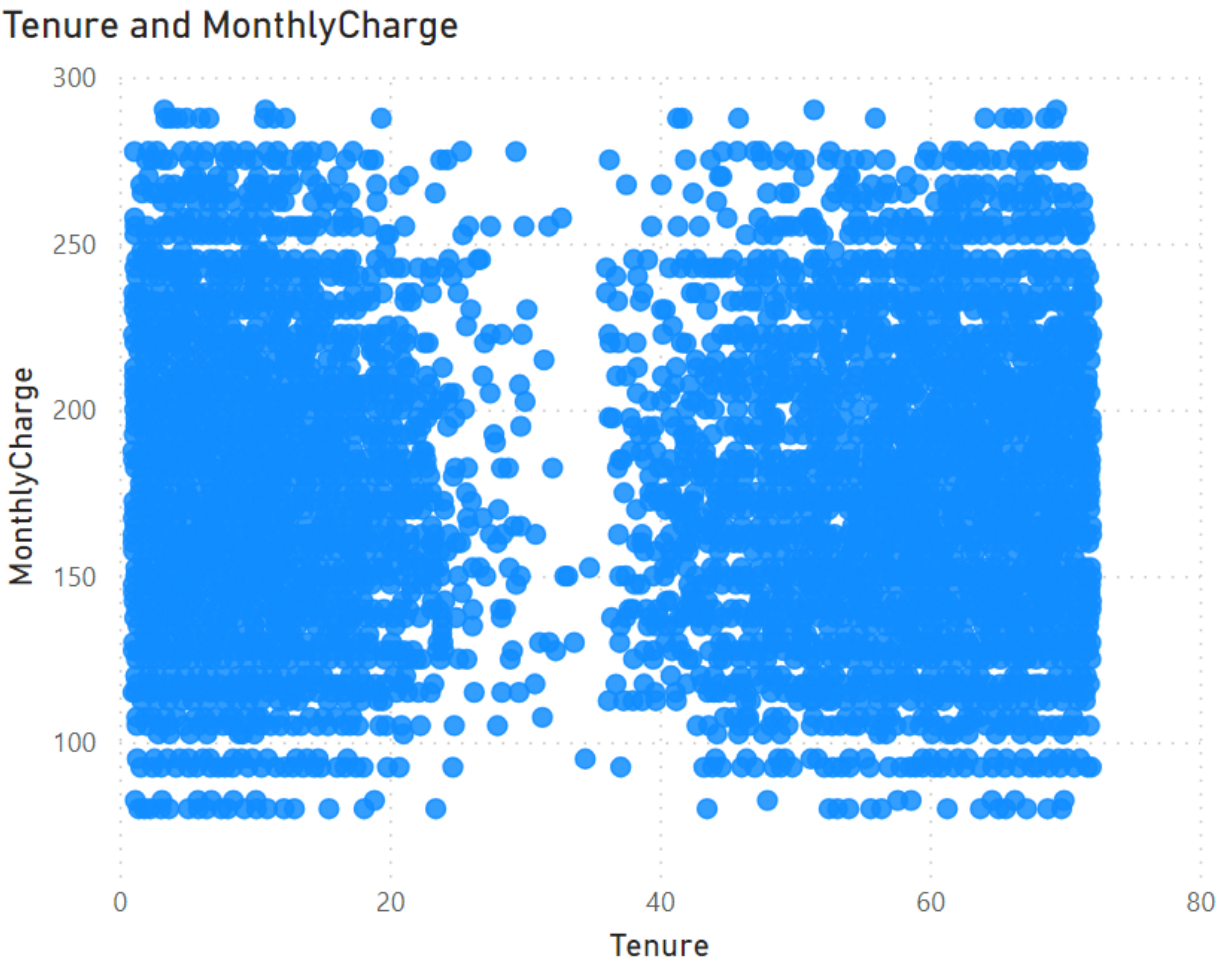
The bivariate graph for gender I created is a stacked bar chart for gender and churn. It is harder to determine if gender has a relationship with churn as the significant change moving from male or female to non-binary results from a small sample of non-binary customers. However, we can say that there is most likely a relationship because moving from the gender categories there is a big jump when it comes to non-binary. My bivariate statistic is the result of the chi-squared analysis that was done earlier comparing the gender to churn. It is not like marital in which all the categories are almost the same value for each category of churn. We can also look further into more bivariate statistics of gender and churn by looking at the crosstab. Here we can see counts of each category of gender by each category of churn.

Gender Churn	Female	Male	Nonbinary
No	3753	3425	172
Yes	1272	1319	59

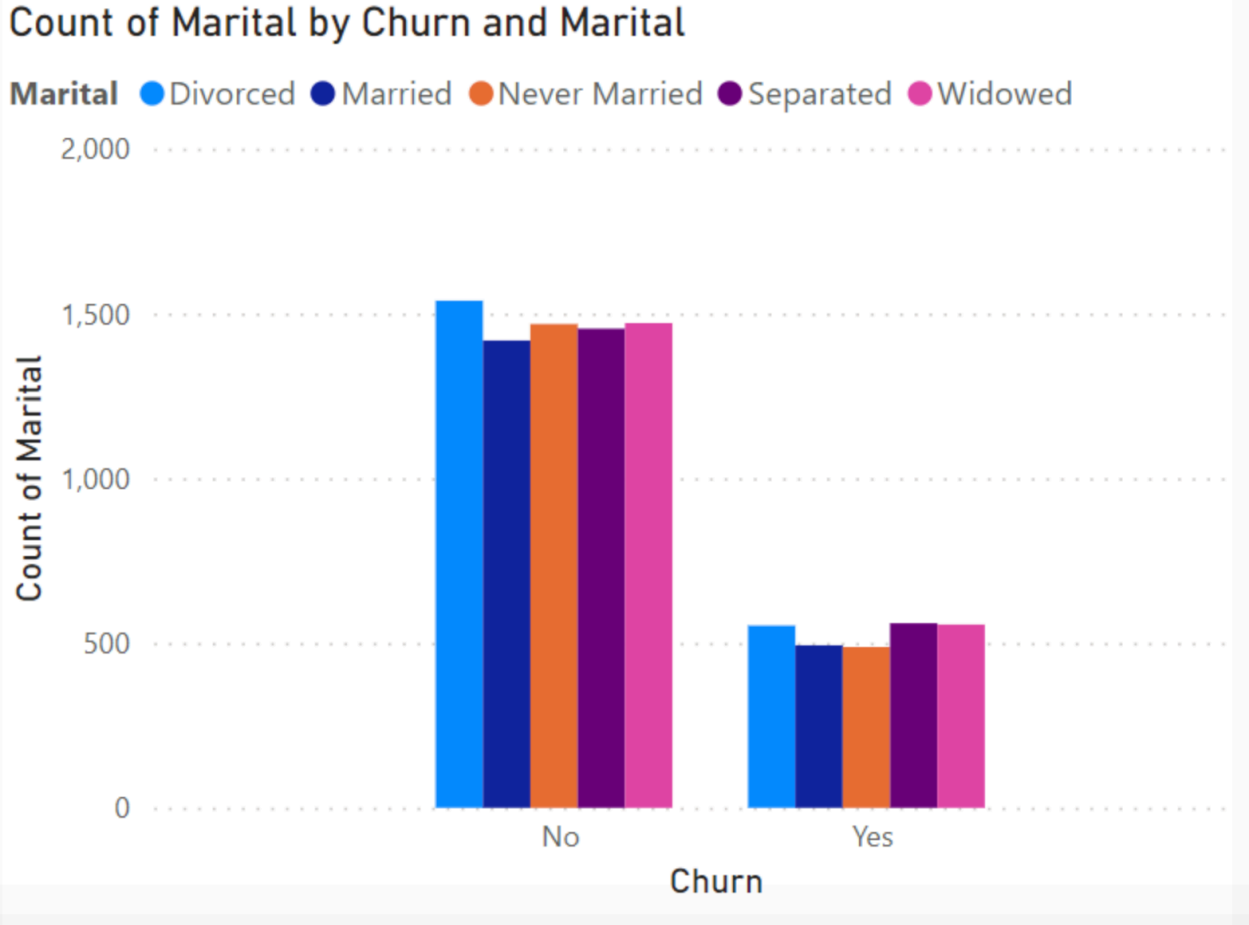
D1: Visual of Findings

Tenure v income

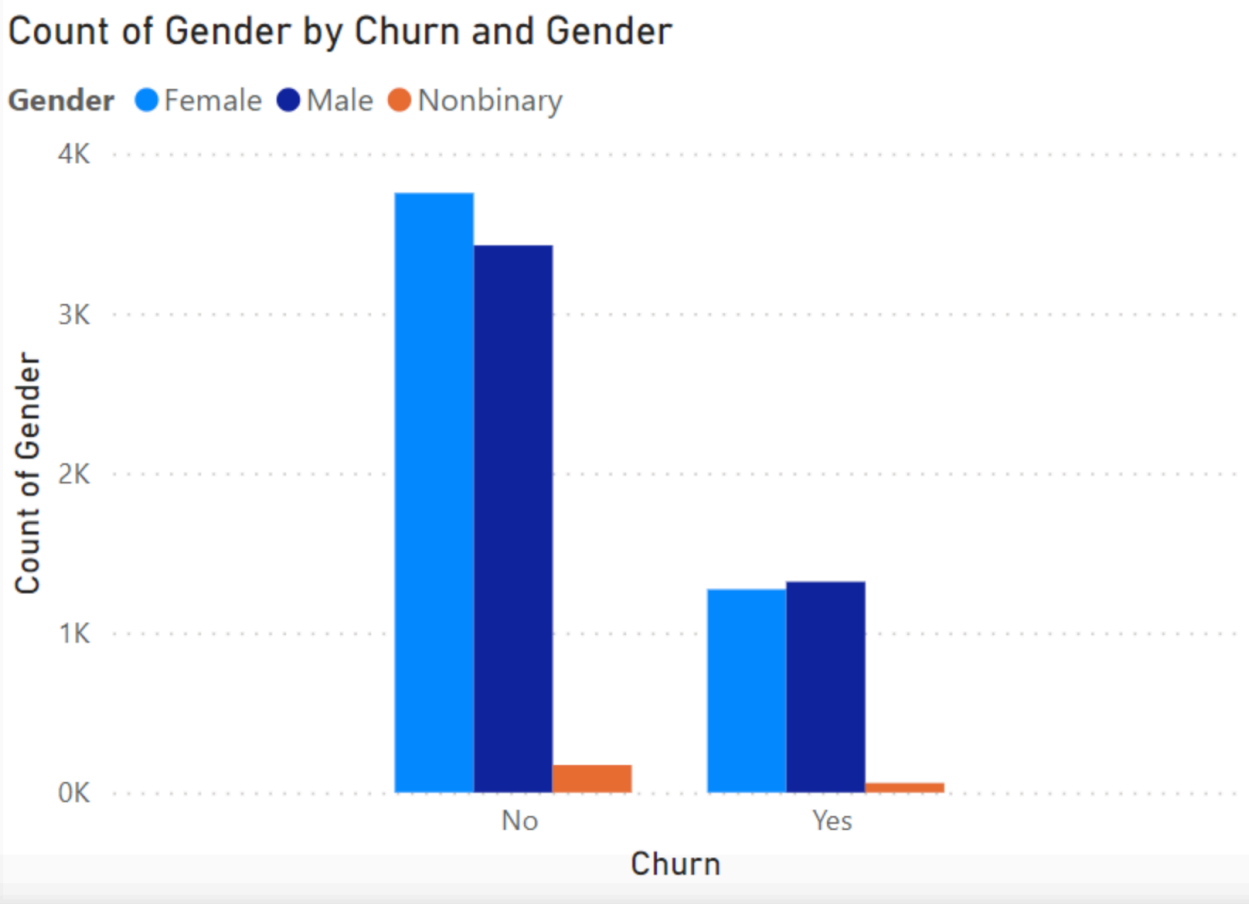
Average tenure vs churn



Marital Status vs Churn (Created in Power BI)



Gender vs churn (Created in Power BI)



E1: Results of Analysis

The results of my analysis are that tenure has a significant relationship with churn. This was tested using both ANOVA analysis. My hypothesis test was:

H_0: there is no significant relationship between the variable and churn

H_a : there is a significant relationship between the variable and churn

. A confidence interval of 95% was used for this test. The output of the ANOVA test between churn and tenure resulted in a p-value below zero, so the alternative hypothesis was accepted, and the null hypothesis was rejected. Income and MonthlyCharge were run on the same ANOVA test to give context to the research question. MonthlyCharge was significantly related to churn, but income was not.

My bivariate analysis can give us even more context to our research question as we can explore income and MonthlyCharge's relationship to tenure. We know from the ANOVA analysis that MonthlyCharge is related to churn, and income is not, so looking at our scatterplots to see if there is an identifiable relationship is a good place to start. We can see from the graphs that there is no obvious linear relationship between tenure and these two variables. We can take from this that it may be more useful if we want to understand more quantitative relationships with churn, we should run full statistical analysis on different variables relation to churn rather than each quantitative variable's relationship to each other. My categorical bivariate visualizations give us more context to some of the categorical variable's relationship to churn. We could further supplement this analysis by doing a chi-squared analysis on the categorical variables to churn. We could also create more visualizations to understand tenure's relationship to these categorical variables.

E2: Limitations of Analysis

The limitations of my analysis are that I did not test every variable. Tenure being the only variable that was the focus of the analysis does not give us a full scope of all the variables that are significant to churn. Even when we use our bivariate visualizations to give context to other variables relationships with churn, we don't fully understand the other variables relationships with churn.

We must also remember that correlation does not assume causation. It's important to note that specifically focusing on a certain variable because it has a positive correlation, for instance increasing the MonthlyCharge through raising prices them may not decrease churn. We have to dig deeper and understand why these relationships exist before making wide ranging assumptions from our data analysis.

E3: Recommended Course of Action

The recommended course of action is to further understand why tenure relationship with churn exist. Answering question like "What business action can we take to increase tenure, and if we increase tenure does that decrease churn?". Also, exploring other variables relationship to tenure. We have bivariate visualizations that show us the gender and marital variables interaction with churn, so we can explore further the impact of tenure on these variables and how that affects churn.

G: Sources for Third-Party Code

No third-party code used

H: Sources

Bevans, R. (2023, June 22). *One-way ANOVA: When and how to use it (with examples)*. Scribbr. https://www.scribbr.com/statistics/one-way-anova/

Hashmi, F. (2020, September 21). *How to visualize data distribution of a categorical variable in python - thinking neuron*. Thinking Neuron - Data Science application to real world problems! https://thinkingneuron.com/how-to-visualize-data-distribution-of-a-categorical-variable-in-python/

Matplotlib labels and title. (n.d.). https://www.w3schools.com/python/matplotlib_labels.asp