PERFORMANCE ASSESSMENT

KAILI HAMILTON

Masters of Science in Data Analytics, Western Governors University

Course: D207 Exploratory Data Analysis

Instructor: Dr. David Gagner

Program Mentor: Krissy Bryant

September, 2023

A1

The research question I will explore in this performance assessment is this: “Is the mean monthly charge for customers who churn significantly different from the mean monthly charge for customers who do not churn?”

I will use a two-sample t-test for means because I am comparing the means for two different groups (Amour Learning, 2019).

A2

Stakeholders in the organization will benefit from an analysis of the data as described in part A1 because having insight into whether customers churn with respect to their monthly charges. For example, if customers are charged too much per month, they may churn and seek business elsewhere. Or, what is the maximum amount to charge customers before they are likely to churn and seek business elsewhere?

A3

The data that are relevant to my research question as described in part A1 are the variables “Churn” and “MonthlyCharge”.

B1

I performed a two-sample t-test on two groups: customers who churn and customers who do not churn. The null hypothesis is that the mean monthly charge for the two groups is the same. The alternative hypothesis is that the mean monthly charge for the two groups is not the same. The alpha value is 0.05 and I will reject the null hypothesis if the p-value is less than alpha (Datacamp, n.d.).

The mean monthly charge for customers who churn is $199.30. The mean monthly charge for customers who do not churn is: $163.01. My code for my calculations is displayed in part B2.

I used Python for my calculations and obtained a t-statistic of 39.29 and a p-value of 0.0. My code for my calculations is displayed in part B2.

Because the p-value is less than alpha, I reject the null hypothesis. So, there is a statistically significant difference in the mean monthly charges for customers who churn and customers who do not churn.

B2

Below are the code snippets from my calculations.

Means of the two groups described in part B1:

```
df_yes['MonthlyCharge'].mean()
```

```
199.29517509886793
```

```
df_no['MonthlyCharge'].mean()
```

```
163.00897252612248
```

T-statistic and p-value for the hypothesis test (DataDaft, 2020):

CALCULATE T-STATISTIC

- alpha = 0.05

```
y_churn = df_yes['MonthlyCharge']
n_churn = df_no['MonthlyCharge']
```

```
stats.ttest_ind(y_churn, n_churn, equal_var=False)
```

```
Ttest_indResult(statistic=39.28778644007045, pvalue=1.7823941678632952e-290)
```

```
# since p < alpha, we reject the null hypothesis
```

Shapiro-Wilkes test for Normality (Datacamp, n.d.:

```
shapiro = stats.shapiro(df_yes.MonthlyCharge)
print(shapiro)
```

```
ShapiroResult(statistic=0.9855512380599976, pvalue=8.825010225524757e-16)
```

```
shapiro = stats.shapiro(df_no.MonthlyCharge)
print(shapiro)
```

```
ShapiroResult(statistic=0.976692259311676, pvalue=5.42239270784402e-33)
```

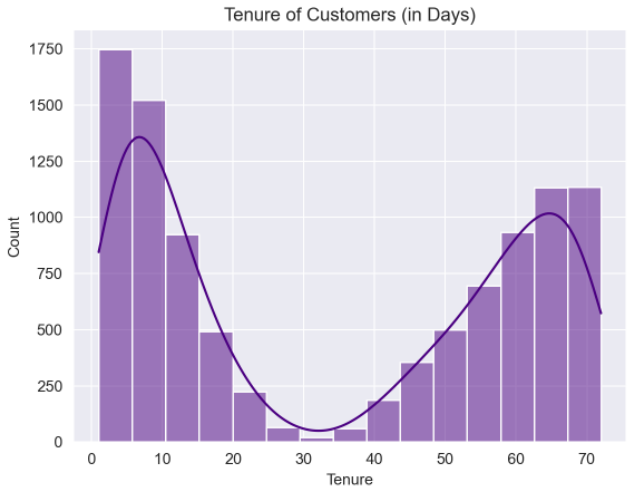
B3

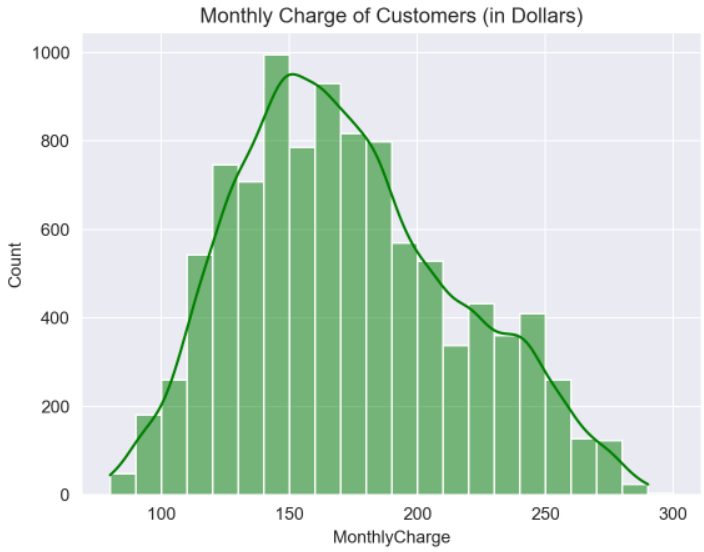
I used a two-sample t-test for means to conduct my analysis because I am comparing the means for two different groups. I want to know whether customers who churn have a significantly different mean monthly charge from customers who do not churn (Amour Learning, 2019).

C1

The distributions of two continuous variables that I considered are “MonthlyCharge” and “Tenure”. Here are the univariate statistics and visualizations (histograms) for these variables (Fessel, 2020):

	Tenure	MonthlyCharge
count	10000.000000	10000.000000
mean	34.526188	172.624816
std	26.443063	42.943094
min	1.000259	79.978860
25%	7.917694	139.979239
50%	35.430507	167.484700
75%	61.479795	200.734725
max	71.999280	290.160419





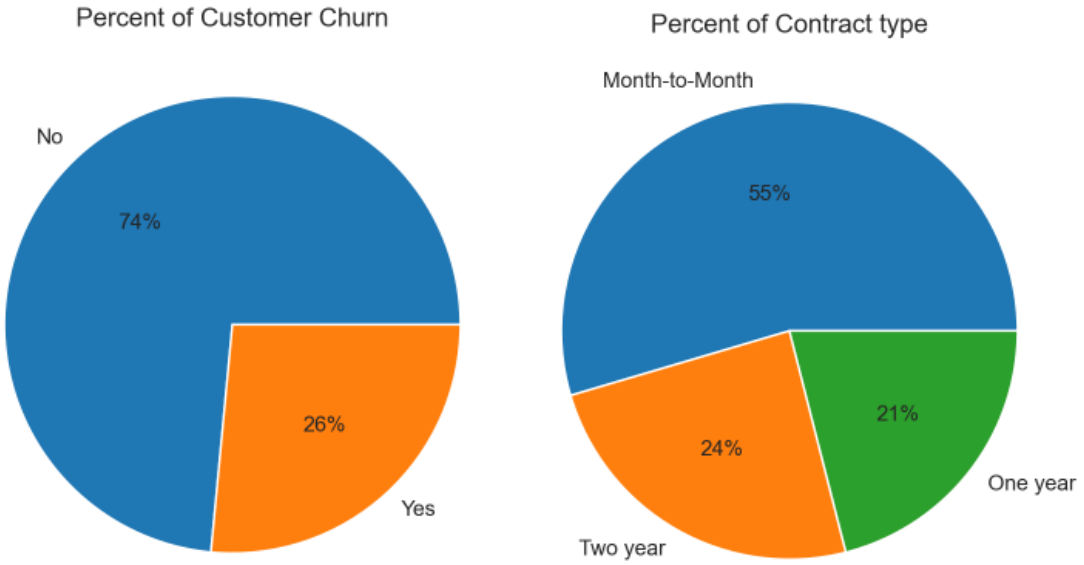
The distributions of two categorical variables that I considered are “Churn” and “Contract”. Here are the univariate statistics and visualizations for these variables (Hunter, 2023):

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

```
Churn
No     7350
Yes    2650
Name: count, dtype: int64
```

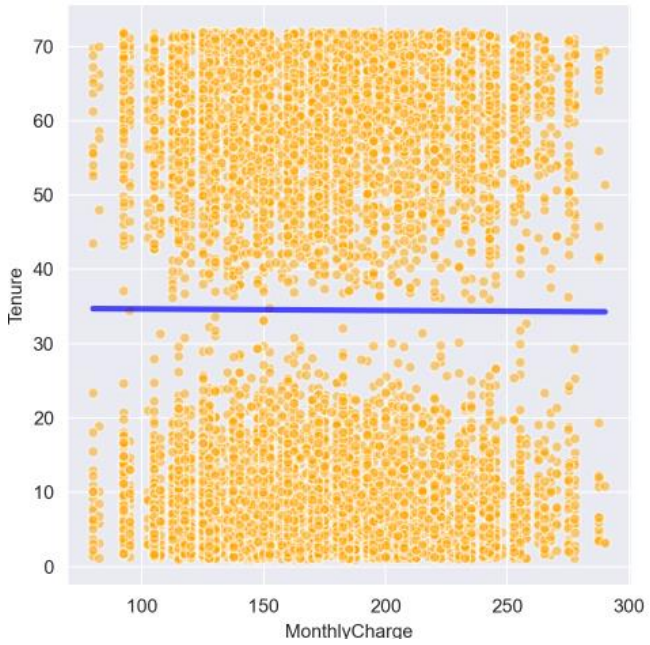
```
df['Contract'].value_counts()
```

```
Contract
Month-to-month    5456
Two Year          2442
One year          2102
Name: count, dtype: int64
```



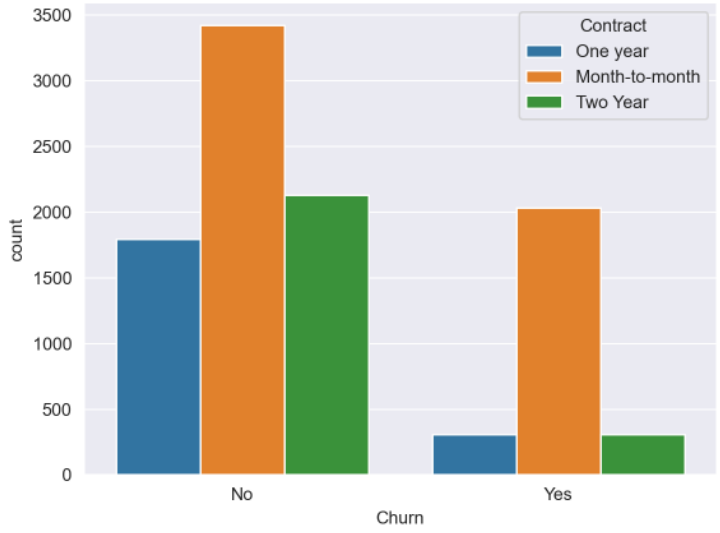
D1

Here I calculated bivariate statistics for the continuous variables “MonthlyCharge” and “Tenure”, with “MonthlyCharge” as the independent variable. The correlation coefficient is $r = -0.003$, meaning there is no correlation between monthly charge and Tenure. The $r^2 = 0.0$, meaning 0% of the variance can be explained between a customer’s monthly charge and their tenure with the company (Fessel, 2021) (Stack Overflow, 2015).



Here I calculated bivariate statistics for the categorical variables “Contract” and “Churn”, seeking whether contract type influences whether customer churn. The table displays the percentage for churn and contract type (Fessel, 2021).

	CONTRACT = MONTH-TO-MONTH	**CONTRACT = ONE YEAR**	**CONTRACT = TWO YEARS**
CHURN = NO	*34.22%*	*17.98%*	*21.33%*
CHURN = YES	*20.34%*	*3.07%*	*3.09%*



E1

I performed a two-sample t-test on two groups: customers who churn and customers who do not churn. The null hypothesis is that the mean monthly charge for the two groups are the same. The alternative hypothesis is that the mean monthly charge for the two groups are not the same. The alpha value is 0.05 and I will reject the null hypothesis if the p-value is less than alpha (Datacamp, n.d.).

The mean monthly charge for customers who churn is $199.30. The mean monthly charge for customers who do not churn is: $163.01. My code for my calculations is displayed in part B2.

I used Python for my calculations and obtained a t-statistic of 39.29 and a p-value of 0.0. My code for my calculations is displayed in part B2.

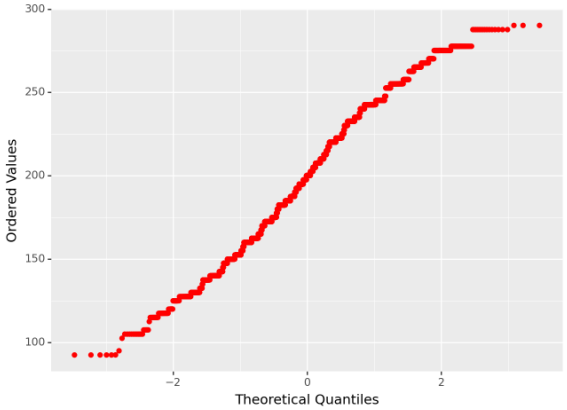
Because the p-value is less than alpha, I reject the null hypothesis. So, there is a statistically significant difference in the mean monthly charges for customers who churn and customers who do not churn (Datacamp, n.d.).

E2

Some limitations of a two-sample t-test of means occur when the assumptions of the test are not met. Assumptions include data values must be independent, data in each group must be obtained via a random sample from the population, and data in each group are normally distributed, data values are continuous (Datacamp, n.d.).

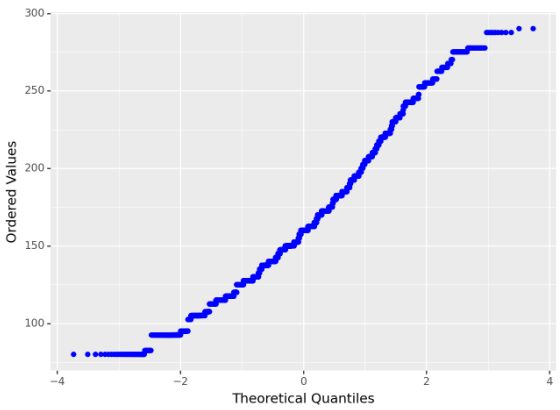
The distributions of customers who churn and customers who do not churn are not normally distributed (see the Q-Q plots and the Shapiro-Wilkes test results below), but the sample sizes are large enough that this assumption can be relaxed (Datacamp n.d.). For customers who churn, n=2650 and for those who do not, n=7350.

Customers Who Churn
Q-Q Plot



ShapiroResult(statistic=0.9855512380599976, pvalue=8.825010225524757e-16)

Customers Who Do Not Churn
Q-Q Plot



ShapiroResult(statistic=0.976692259311676, pvalue=5.42239270784402e-33)

E3

Customers who churn have a mean monthly charge of $199.30. Customers who do not churn have a mean monthly charge of $163.01. The results of our hypothesis test (described above) indicate that the mean monthly charge for these two groups are significantly different.

I recommend a course of action to consider ways of lowering the mean monthly charge. We can see that customers who are not churning pay an average of $36 less per month than customers who are churning. If a customer is with the company for more than a year, as an example, can their monthly charge be discounted? Could a tenured customer receive other perks to aid in lowering monthly charges? Are there discounts that can be offered for bundling packages? Are customers paying for services they actually use? If not, customer service could reach out to customers inquiring as to whether they'd like to discontinue particular services that they aren't using, thus lowering their monthly charges.

F

A Panopto video recording including a demonstration of the functionality of my code used for my analysis is provided with the submission of my Performance Assessment.

G

Amour Learning (2019, September 5). How To Know Which Statistical Test To Use For Hypothesis Testing [Video]. YouTube. https://www.youtube.com/watch?v=ChLO7wwt7h0

Datacamp. (n.d.). Foundations of Probability in Python. Datacamp. https://campus.datacamp.com/courses/foundations-of-probability-in-python/probability-meets-statistics?ex=1

DataDaft (2020, December 15). Python for Data Analytics: Hypothesis Testing and T-Tests [Video]. https://www.jmp.com/en_us/statistics-knowledge-portal/t-test/two-sample-t-test.html

Fessel, K. (2020, December 7). Seaborn histplot | How to make a Seaborn histogram plot with Python code [Video]. YouTube. https://www.youtube.com/watch?v=Bjz00ygERxY

Fessel, K. (2021, April 12). Seaborn lmplot | Comparing lmplot vs regplot and using the Seaborn lmplot hue and FacetGrid [Video]. YouTube. https://www.youtube.com/watch?v=-_g_z-L-6tc

Fessel, K. (2021, April 26). Seaborn countplot | What is the countplot? | Seaborn countplot vs barplot [Video]. YouTube. https://www.youtube.com/watch?v=8U5h3EJuu8M

How to make seaborn regplot partially see through (alpha). (2015). Stack Overflow. Retrieved from https://stackoverflow.com/questions/33005872/how-to-make-seaborn-regplot-partially-see-through-alpha

Hunter, J. et al. (2023). Pie Charts. Mathplotlib. https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

H

I acknowledged sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I

Professional communication is demonstrated in the content and presentation of my Performance Assessment.