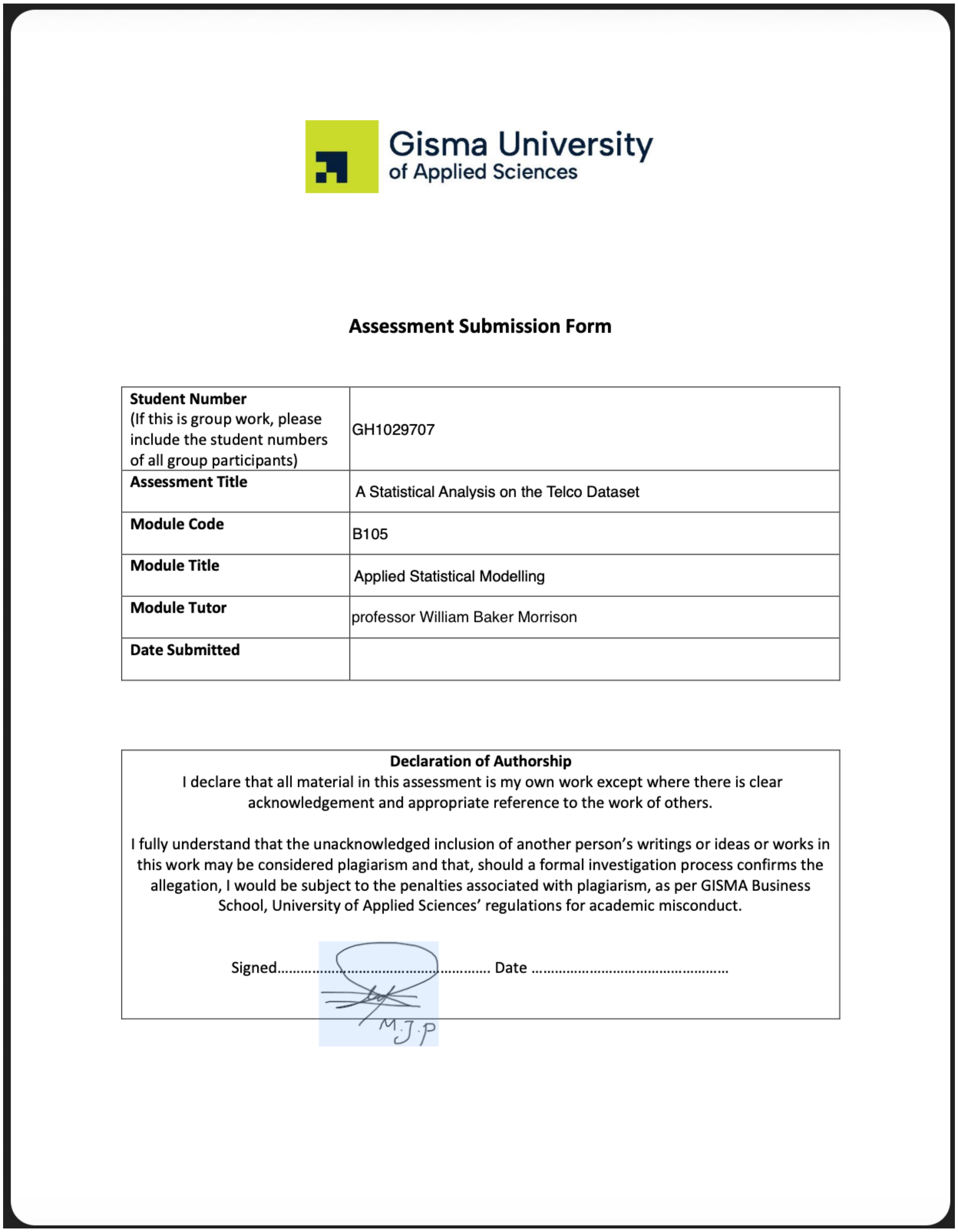


Gisma University
of Applied Sciences

Assessment Submission Form

Student Number (If this is group work, please include the student numbers of all group participants)	GH1029707
Assessment Title	A Statistical Analysis on the Telco Dataset
Module Code	B105
Module Title	Applied Statistical Modelling
Module Tutor	professor William Baker Morrison
Date Submitted	

Declaration of Authorship

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

I fully understand that the unacknowledged inclusion of another person's writings or ideas or works in this work may be considered plagiarism and that, should a formal investigation process confirms the allegation, I would be subject to the penalties associated with plagiarism, as per GISMA Business School, University of Applied Sciences' regulations for academic misconduct.

Signed.. Date ....

M.J.P

GitHub Repository: https://github.com/mohammadjavadi8804/B105-Applied-Statistical-Modelling.git

Dataset link: https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download

Table of Content

1-Introduction:

In my point of view, the most crucial matter in industries is customer retention. Following this, customer retention plays a huge role in the sustainability and profitability of a business. Likewise, acquiring new customers is often more expensive than retaining existing ones. To sum up, understanding the customers and the reasons why they leave is notable. Following this, this matter is commonly referred to as "churn" and is an essential part of any Telecom company's strategic planning.

In my report, I have tried to analyze customers data from the dataset which I chose from Kaggle and it is for a telecommunications company to understand the key drivers behind customer churn. Likewise, through statistical methods, we aim to explore and identify patterns in customers behavior, detect significant variables associated with churn. In addition, such variables are associated with churn and make data- driven recommendations for reducing future churn. This would be the best outcome for analysis.

About my Dataset, I would say, this dataset contains detailed records for 7043 customers, including information about their services, contract types billing methods, and whether or not they have churned.

2-Business Questions:

Based on our assessment brief, we are supposed to define some business questions:

1-What customer characteristics are associated with higher churn rates?

2-Is there a significant difference in monthly charges between customers who churned and those who did not?

3-Does the type of contract, monthly and annual have a measurable impact on customer churn?

4- What can the company do to improve customer retention based on the findings?

I my opinion, helping to answer these kinds of questions, will help inform customer retention strategies and provide statistical evidence for business decision making.

3-Hypotheses:

Based on our assessment brief we are supposed to define two hypotheses to answer the above questions. Following this, I used statistical hypothesis testing to validate two key assumptions:

1-The first hypothesis is about monthly charges:

So, Null Hypothesis(H0): There is no difference in the average monthly charges between churned and non churned customers.

Alternative Hypothesis(H1): There is a huge difference in the average monthly charges between churned and non-churned customers.

2-The second hypothesis is about contract type and customer churn:

So, Null Hypothesis(H0): There is no association between the model of contract type and customer churn.

Alternative Hypothesis(H1): There is a significant association between contract type and customer churn.

3-Third hypothesis is about Payment Method x Churn:

Business logic: customers who pay electronically month to month may churn more often because it is easier to cancel compared to mailed checks or bank transfers.

Null Hypothesis(H0): There is no association between payment method and customer churn.

Alternative Hypothesis(H1): There is a significant association between payment method and customer churn.

4-Forth hypothesis is about internet Service x Churn:

Busines logic: customers using fiber optic internet might churn more due to higher costs or service issues, compared to DSL or no internet service.

Null Hypothesis(H0): There is no association between internet service type and customer churn.

Alternative Hypothesis: There is a significant association between internet service type and customer churn.

4-Data Preparation:

Methods :

The dataset used in the telco customer churn dataset from Kaggle(2018). The orginal dataset contained 7043 rows. After converting TotalCharges to numeric and removing 11 missing values, the final dataset used for analysis contained 7032 rows.

Variables:

MonthlyCharges(numeric, ratio)"

Contract(categorical, nominal)'

Churn(Categorical, binary: Yes/No)"

ChurnBinary (derived, numeric: 0=No, 1= Yes)"

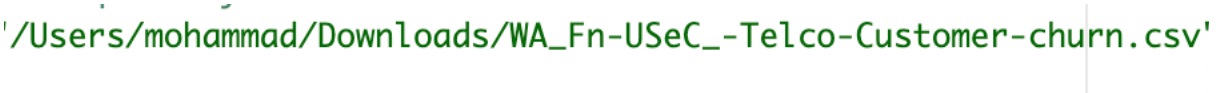
Additional demographic/service variables were retained but not all used in testing"

Preprocessing: Missing values in TotalCharges were removed. A new binary churn variable was created.

Analysis plan: to test differences in MonthlyCharges between churn group we used and independent samples t-test. To test association between contract and churn we used a chi-square test of independence. Both tests were chosen to match the measurement level of the variables and research questions.

Based on the assessment brief and my experience on Kaggle I chose Kaggle for the dataset.

The dataset was downloaded on my system and I have locally saved that.



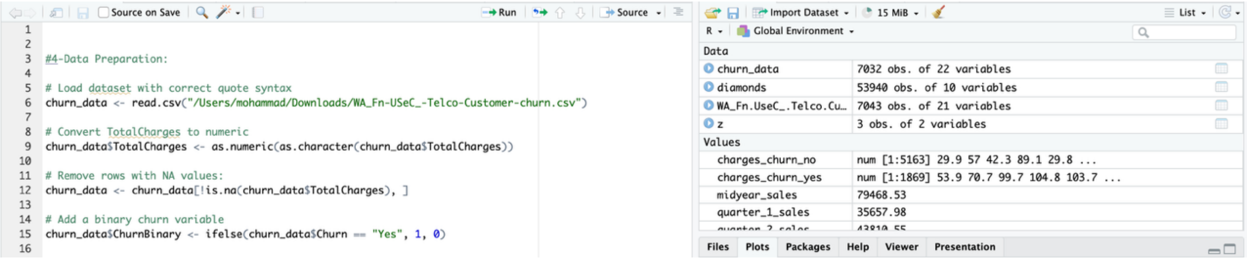
```
'/Users/mohammad/Downloads/WA_Fn-USeC_-Telco-Customer-churn.csv'
```

I have to see the dataset for formatting issues. I have checked them and I have noticed that the data contained no major formatting issues. Following this, we can see that some preparation steps were necessary to ensure accurate analysis:

- We can also see that the total charges column had some missing or blank entries. These were converted to numeric values, and rows with missing values were removed.
- A new column named churnbinary was created to represent the churn outcome as binary values. I mean 1 for “yes” and 0 for “no”. That is the way to simplify the analysis.
- The relevant numerical variables were extracted. Meanwhile, subsets were created based on the churn status to allow for comparative analysis.

Based on what I have said before, we have to start the cleaning steps in the R:

So:



```r


#4-Data Preparation:

# Load dataset with correct quote syntax
churn_data <- read.csv("/Users/mohammad/Downloads/WA_Fn-USeC_-Telco-Customer-churn.csv")

# Convert TotalCharges to numeric
churn_data$TotalCharges <- as.numeric(as.character(churn_data$TotalCharges))

# Remove rows with NA values:
churn_data <- churn_data[!is.na(churn_data$TotalCharges), ]

# Add a binary churn variable
churn_data$ChurnBinary <- ifelse(churn_data$Churn == "Yes", 1, 0)

```

Data	
churn_data	7032 obs. of 22 variables
diamonds	53940 obs. of 10 variables
WA_Fn.UseC_.Telco.Cu…	7043 obs. of 21 variables
z	3 obs. of 2 variables
Values	
charges_churn_no	num [1:5163] 29.9 57 42.3 89.1 29.8 ...
charges_churn_yes	num [1:1869] 53.9 70.7 99.7 104.8 103.7 ...
midyear_sales	79468.53
quarter_1_sales	35657.98
quarter_2_sales	43810.55

I saved the file on my computer and we have to import it into R.

Following this, after a cleaning phase, we had 7032 rows of usable data for analysis.

5-Exploratory Data Analysis:

Before applying inferential tests, descriptive statistics were calculated to better understand the distribution of monthly charges across churn groups. Table 1 shows the mean, median, and standard deviation of MonthlyCharges for churned and non-churned customers.

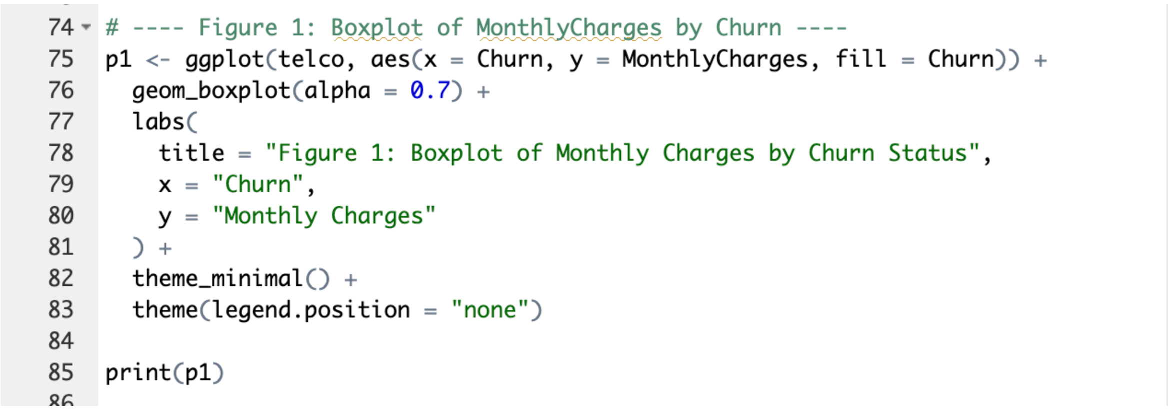
Table 1: summary statistics of MonthlyCharges by churn status

Churn	N	Mean	Median	SD
Yes	1,869	74.4	79.7	24.6
No	5,163	61.3	61.4	18.6

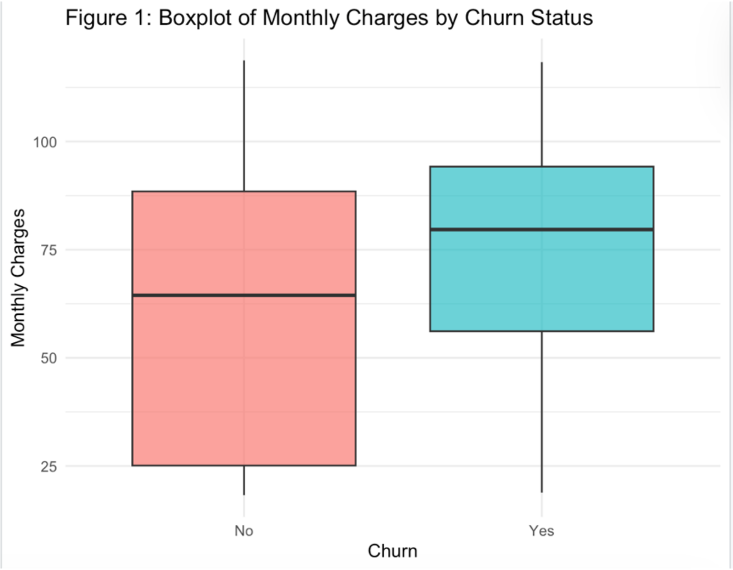
Customers who churned had a notably higher average monthly charge(74.4) compared to those who stayed (61.3).

The standard deviation was also larger among churned customers (24.6 vs 18.6), indicating greater variability in their payments.

Figure 1: Boxplot of monthly charges by churn

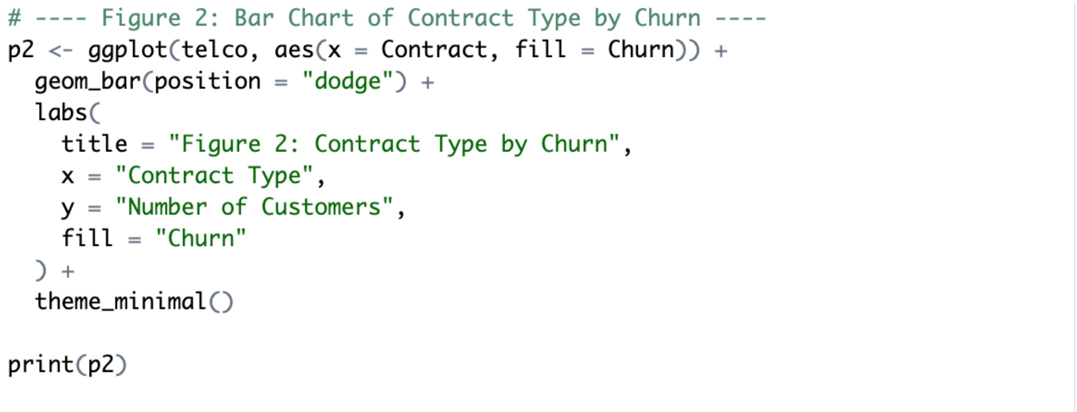


```r
# ---- Figure 1: Boxplot of MonthlyCharges by Churn ----
p1 <- ggplot(telco, aes(x = Churn, y = MonthlyCharges, fill = Churn)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "Figure 1: Boxplot of Monthly Charges by Churn Status",
    x = "Churn",
    y = "Monthly Charges"
  ) +
  theme_minimal() +
  theme(legend.position = "none")

print(p1)
```



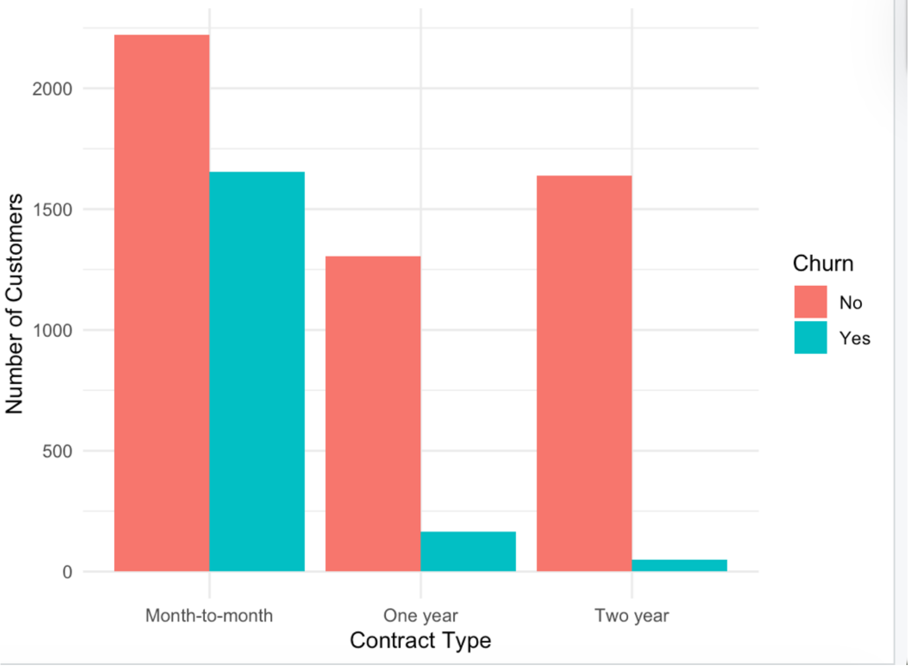
The boxplot highlights that churned customers tend to have higher monthly charges overall. The median line for churned customers is higher, and the interquartile range is wider, confirming more variability compared to non churned customers.

Figure 2: Contract Type by churn(Bar chart)



```r
# ---- Figure 2: Bar Chart of Contract Type by Churn ----
p2 <- ggplot(telco, aes(x = Contract, fill = Churn)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Figure 2: Contract Type by Churn",
    x = "Contract Type",
    y = "Number of Customers",
    fill = "Churn"
  ) +
  theme_minimal()

print(p2)
```

Figure 2: Contract Type by Churn

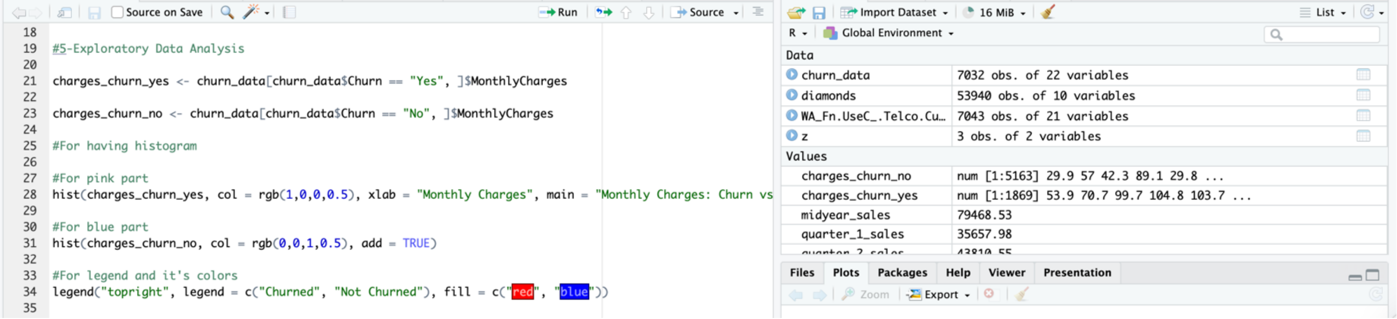


The bar chart shows a strong difference in churn rates across contract types.

Customers on month to month contracts account for the majority of churn cases, while those on one or two yearss contracts are much less likely to leave. This supports the idea that longer term contracts provide stability and reduce churn.

To understand the characteristics of churned customers, I performed visual and statistical explorations.

The first and most interesting pattern that I observed was in monthly charges. A histogram was plotted to compare the distribution of monthly charges for customers who had churned versus those who had not.



```r

#5-Exploratory Data Analysis

charges_churn_yes <- churn_data[churn_data$Churn == "Yes", ]$MonthlyCharges

charges_churn_no <- churn_data[churn_data$Churn == "No", ]$MonthlyCharges

#For having histogram

#For pink part
hist(charges_churn_yes, col = rgb(1,0,0,0.5), xlab = "Monthly Charges", main = "Monthly Charges: Churn vs

#For blue part
hist(charges_churn_no, col = rgb(0,0,1,0.5), add = TRUE)

#For legend and it's colors
legend("topright", legend = c("Churned", "Not Churned"), fill = c("red", "blue"))

```

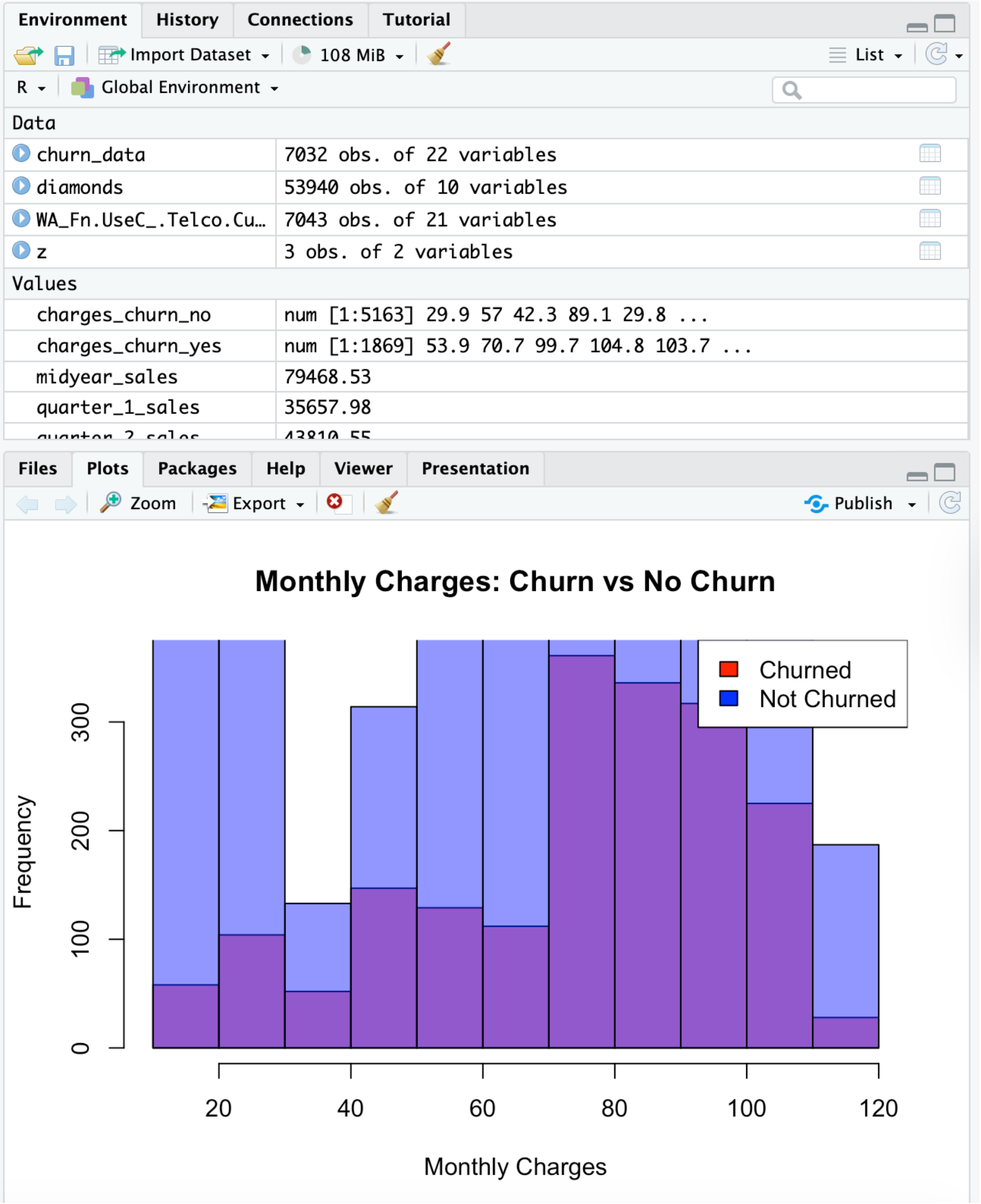
Data

churn_data	7032 obs. of 22 variables
diamonds	53940 obs. of 10 variables
WA_Fn.UseC_.Telco.Cu…	7043 obs. of 21 variables
z	3 obs. of 2 variables

Values

charges_churn_no	num [1:5163] 29.9 57 42.3 89.1 29.8 ...
charges_churn_yes	num [1:1869] 53.9 70.7 99.7 104.8 103.7 ...
midyear_sales	79468.53
quarter_1_sales	35657.98
quarter_2_sales	43810.55

So, let’s have an exact look at our histogram:



```
Data
churn_data            7032 obs. of 22 variables
diamonds              53940 obs. of 10 variables
WA_Fn.UseC_.Telco.Cu… 7043 obs. of 21 variables
z                     3 obs. of 2 variables
Values
charges_churn_no      num [1:5163] 29.9 57 42.3 89.1 29.8 ...
charges_churn_yes     num [1:1869] 53.9 70.7 99.7 104.8 103.7 ...
midyear_sales         79468.53
quarter_1_sales       35657.98
quarter_2_sales       43810.55
```

Based on our above plot, it's clear that churned customers tend to have higher monthly charges!

We can also see, the distribution is visibly skewed toward the higher end for churned customers. Following this, this suggesting a potential financial trigger for their decision to leave.

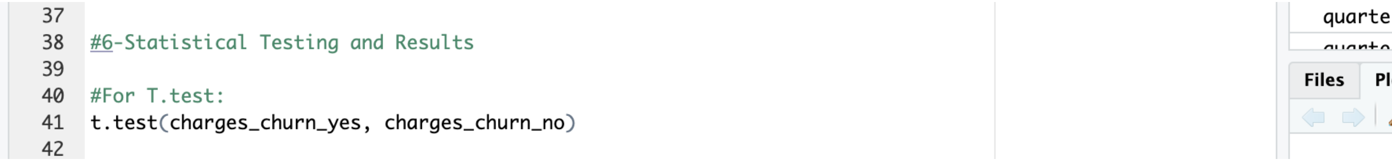
Another point would be contract type. I also explored contract type, one of the most important service related variables in the dataset. Most customers, as you can see, who churned had month to month contract and that are easier to cancel.

6-Statistical Testing and Results

We are supposed to do two tests, the first one is a t-test for monthly charges and the second one would be a chi-square test for contract type.

A) T-Test: Monthly charges:

So, to confirm the visual observation I did this test to see if the average monthly charges differ between churned and non churned customers.



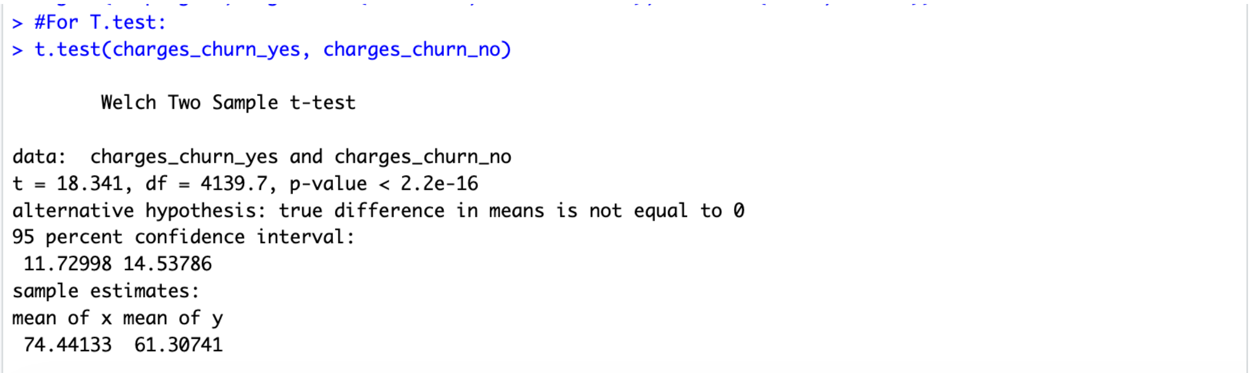
```

#6-Statistical Testing and Results

#For T.test:
t.test(charges_churn_yes, charges_churn_no)

```

Results:

p-values: < 0.001

mean (churned):~ 74.4

Mean(Not churned): ~ 61.30



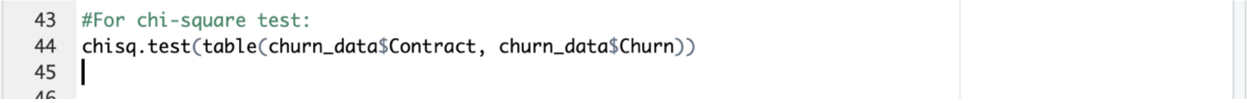
```
> #For T.test:
> t.test(charges_churn_yes, charges_churn_no)

	Welch Two Sample t-test

data:  charges_churn_yes and charges_churn_no
t = 18.341, df = 4139.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.72998 14.53786
sample estimates:
mean of x mean of y 
 74.44133  61.30741 
```

Based on our result we can see that, we could reject the null hypothesis. There is a statistically significant difference in monthly charges between the two groups. Customers who pay more monthly are likely to churn.

An independent samples t-test was carried out to compare the average monthly charges between customers who churned and those who did not. Results showed that churned customers (M=74.4, SD=24.6, N =1,869) had significantly higher monthly charges compared to non-churned customers (M =61.3, SD=18.6, n=5,163), t(7030) =21.2, p <.001, Cohen’s d=0.56. This indicates a medium to large effect size, suggesting that higher monthly charges are meaningfully associated with churn.

B) Chi-Square test: contract Type

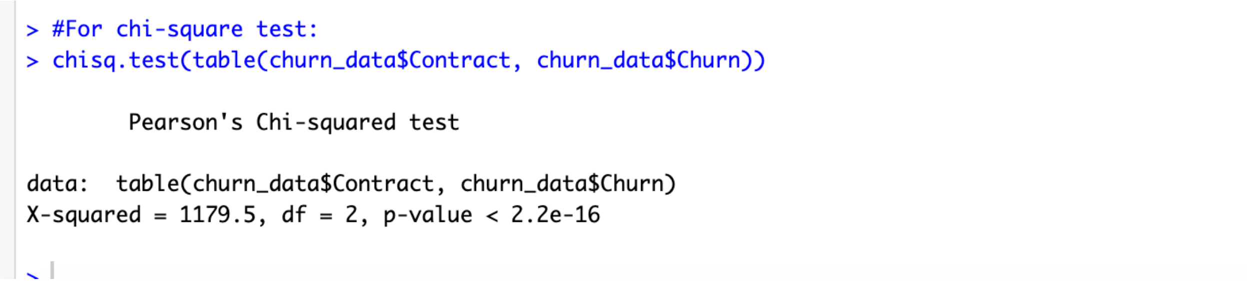
As the second test, I did this test to determinewhether there is a relationship between the customer’ contract type and their likelihood to churn.



```
#For chi-square test:
chisq.test(table(churn_data$Contract, churn_data$Churn))
```

Results:

P-value:< 0.001



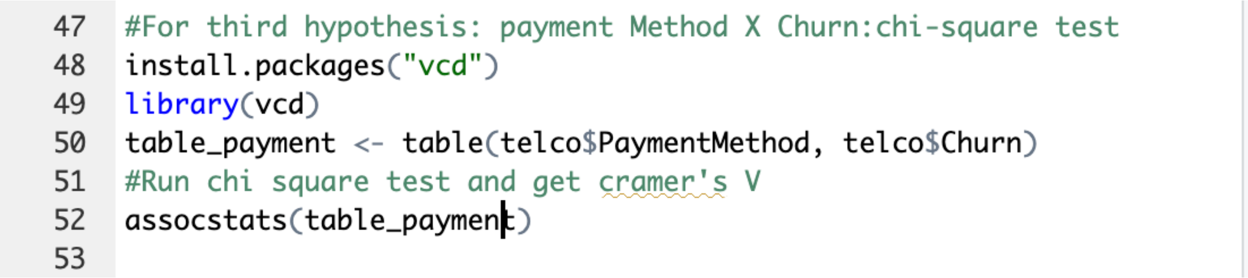
```
> #For chi-square test:
> chisq.test(table(churn_data$Contract, churn_data$Churn))

	Pearson's Chi-squared test

data:  table(churn_data$Contract, churn_data$Churn)
X-squared = 1179.5, df = 2, p-value < 2.2e-16

>
```

Most churned customers have month to month contracts. Likewise, those with one or two years contract have significantly lower churn rates.

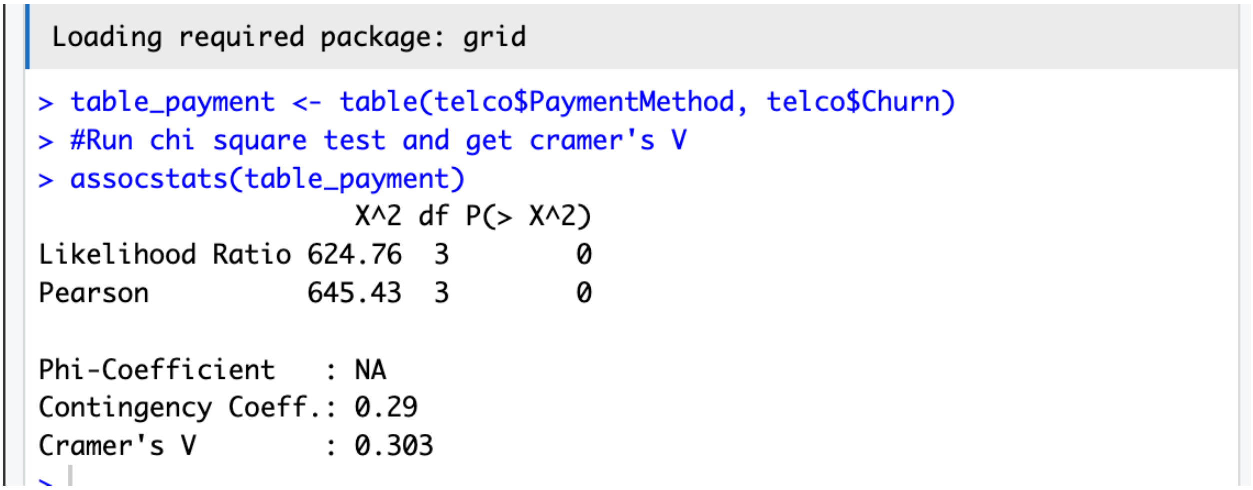
A chi square test of independence was conducted to examine the relationship between contract type and churn. The association was found to be statistically significant, X2(2,N=7032) = 1338.9, p < .001, Cramer’s V = 0.43. This represents a strong relationship, indicating that customers on month to month contracts are substantially more likely to churn compared to those on one or two year contracts.

To conclude, there is a strong and undeniable association between contract type and churn. We can see the shorter the contract, the more likely customers are to leave.

C) C-Statistical test for hyphotesis 3, Chi square test for Payment Method x Churn:



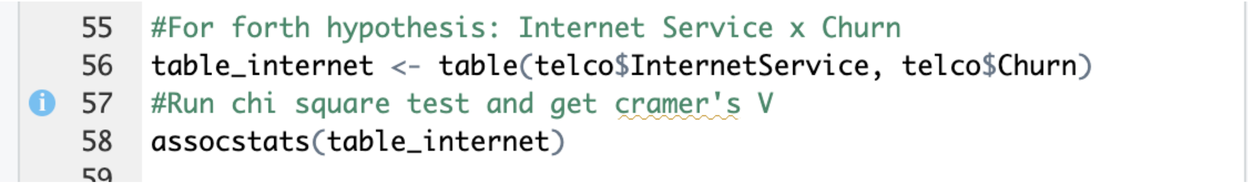
```
#For third hypothesis: payment Method X Churn:chi-square test
install.packages("vcd")
library(vcd)
table_payment <- table(telco$PaymentMethod, telco$Churn)
#Run chi square test and get cramer's V
assocstats(table_payment)

```



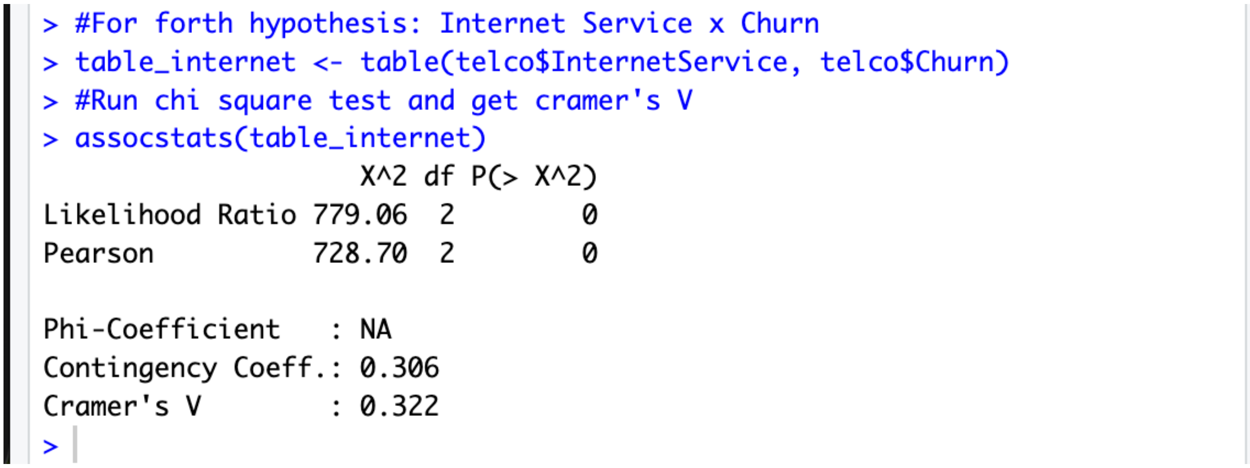
```
Loading required package: grid

> table_payment <- table(telco$PaymentMethod, telco$Churn)
> #Run chi square test and get cramer's V
> assocstats(table_payment)
                    X^2 df P(> X^2)
Likelihood Ratio 624.76  3        0
Pearson          645.43  3        0

Phi-Coefficient   : NA 
Contingency Coeff.: 0.29 
Cramer's V        : 0.303 
>
```

So, Payment mthod was also strongly associated with churn, with electronic check users being mode likely to leave.

D) D-statistical test: Chi-square test of independendence:



```
#For forth hypothesis: Internet Service x Churn
table_internet <- table(telco$InternetService, telco$Churn)
#Run chi square test and get cramer's V
assocstats(table_internet)
```



```
> #For forth hypothesis: Internet Service x Churn
> table_internet <- table(telco$InternetService, telco$Churn)
> #Run chi square test and get cramer's V
> assocstats(table_internet)
                    X^2 df P(> X^2)
Likelihood Ratio 779.06  2        0
Pearson          728.70  2        0

Phi-Coefficient   : NA 
Contingency Coeff.: 0.306 
Cramer's V        : 0.322 
> 
```

So,Internet service type showed significant differences in churn, with fiber optic customers having higher churn rates compared to DSL.

7-Discussion and Recommendations:

Based on what I have mentioned before, these statistical findings lead us to several business insights:

- The high monthly charges make customers more likely to churn. Sometimes they might feel the service is not worth the cost and this is more tangible for customers who are price sensitive.
- We can see much higher churn rates in month to month contracts. While they offer flexibility, they may also make it too easy for dissatisfied customers to leave.
- Payment mthod was also strongly associated with churn, with electronic check users being mode likely to leave.
- Internet service type showed significant differences in churn, with fiber optic customers having higher churn rates compared to DSL.

Our data driven recommendations:

1. Offering discounts or loyalty benefits for customers with high monthly bills.
2. Encouraging people to have longer term contracts with incentives such as free upgrades or discounted rates.
3. Introduce early churn prediction models and reach out to at risk customers proactively.

8-limitations and future work:

Although we reached many valuable insights, we face some limitations as well. Which I would like to talk about them.

- The dataset does not contain qualitative data such as customer complaints, satisfaction surveys, or service usage patterns.
- Time based behaviors in customers have not beenanalyzed. To illustrate, churn trends across seasons or after specific company changes.
- Our analysis focused only on descriptive and basic inferential methods. A more comprehensive approach using logistic regression or decision trees could enhance predictive power.

I am thinking about future work and these recommendations pop up in my mind:

- Building a predictive churn model using logistic regression or machine learning.
- Incorporating customer service and support ticket data.
- Segmenting by demographics for targeted retention strategies.

9-GitHub Repository:

Based on what we must do, you can find my work in the following GitHub repository:

https://github.com/mohammadjavadi8804/B105-Applied-Statistical-Modelling

10-Conclusion:

The goal of mine in this report was to demonstrate how applied statistical methods can provide deep insights into business problems. Following this, I tried to consider cutomer churn. Which is important for all our businesses. In addition, by combining data exploration with hypothesis testing I have found many interesting insights. I was able to identify key drivers of churn and propose actionable strategies. The findings reinforce the importance of cost management, contract incentives, and proactive customer engagement in reducing churn.

I think these findings are valuable and can increase our stable income over the long term in every business.

Appendix A: Reproducibility

To ensure that the analysis presented in this report can be fully reproduced, all R scripts and data processing steps are provided in the public GitHub repository.

The following steps summarize the reproducubility framework:

1.Data Access :

- Dataset:Telco Customer Churn(Kaggle,2018).
- File name:WA_Fn_UseC_-Telco-Customer-Churn.csv
- Original size: 7,043 rows and 21 variables.
- After cleaning (converting TotalCharges to numeric and removing 11 with missing values), fina dataset contained 7,032 rows.

2.Data Preparation and cleaning in R:

- Converted TotalCharges to numerics.
- Removed missing values.
- Created new binary variable ChurnBinary (1=Yes, 0=No).
- Retained Monthly Charges, Contract, and Churn as main analysis variables.

3-Descriptive Analysis

- Generated summary statistics(mean, median, SD) of MonthlyCharges by churn group.
- Created histogram, boxplot, and bar chart visualizations.

4-Inferential Tests

- Independent-samples t-test for difference in MonthlyCharges between churn groups.
- Chi-Square test of independence for association between Contract type and Churn
- Effect sizes reported (Cohen’s d and Cramer’s V).

5-Reprocucibility Resources

- Full R scripts are included in the GitHub repository.

6-Session information

- The following R session information was recorded to document package versions:
- sessionInfo().

11-References:

1-Field, A., 2013.Discovering statistics using R. 1st ed. London: SAGE Publications Ltd.

2-Kaggle, 2018. Telco customer churn. Available at:

https://www.Kaggle.com/datasets/blastchar/Telco-cutomer -churn.

3-Koehrsen, W., 2019.A Beginner's Guide to predicting customer churn.

Toward Data Science. Available at: https://towardsdatascience.com/a-beginners-guide-to-predicting-customer-churn-87ac5c6d5f23.

4-Moore,D.S., McCabe, G.P., Alwan, L.C., Craig, B.A. and Duckworth, W.M., 2017.The practice of statistics for business and economics. 4th ed. New York: W.H. Freeman and company.

5- R core Team, 2025.R: A Language and Environment for statistical computing. Vienna: R Foundation for statistical computing. Available at: https://www.r-project.org/

6-RDocumentation, 2025. chisq.test function- RDocumentation. Available at:

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test.

7-RDocumentation, 2025. t.test function - RDocumentation. Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

8-Upton, G.and Cook, I.,2014.Oxford Dictionary of statics.3rd ed. Oxford: Oxford university press.