D207 Exploratory Data Analysis

Scott Babcock WGU - MS, Data Analytics

Created: October 22 2024

R Packages Used in Analysis

```
# Load libraries that will be used
library(dplyr)
library(ggplot2)
library(ggpubr)
library(naniar)
library(broom)
```

Turn off Scientific Notation

```
# Turn of displaying numbers in Scientific Notation [In-text citation: (Bobbitt, 2021)]
options(scipen = 999)
```

Initial Data Load

```
# Load Churn data set file
churn_df <- read.csv("churn_clean.csv")

# Check for missing data [In-text citation: (Tierney, n.d.)]
miss_var_summary(churn_df)
```

```
## # A tibble: 50 x 3
##    variable    n_miss pct_miss
##    <chr>        <int>    <num>
##  1 CaseOrder        0        0
##  2 Customer_id      0        0
##  3 Interaction      0        0
##  4 UID              0        0
##  5 City             0        0
##  6 State            0        0
##  7 County           0        0
##  8 Zip              0        0
##  9 Lat              0        0
## 10 Lng              0        0
## # i 40 more rows
```

```
# Create data frame for variables used in statistical testing
churn_test <- churn_df %>%
    select(Churn, Income)
```

A1, Research Question

Do the income levels of customers who have discontinued service in the last month differ from active customers?

A2, Benefit to the Organization

In the telecommunications industry, it is vital to retain customers. The cost of acquiring a new customer dramatically exceeds that of retaining an existing one. By analyzing the available customer data and identifying differences amongst

groups, the organization can gain valuable insights into whether certain factors make it more likely that a customer will discontinue service. Suppose there is a statistically significant difference in the mean incomes of customers who currently have services versus those who have canceled. In that case, the organization can identify customers at risk of discontinuing service and be more thoughtful about which new customers they target.

A3, Data Used

Only two fields are needed to answer the question regarding the income levels of current and former customers. The Churn field, a Yes/No categorical data type, tracks whether a customer has canceled their services in the last month. The Income field, a numeric data type, provides the customer's annual income.

B1, Statistical Technique

The R programming language was used to test the research question. The technique chosen was a two-sample t-test. The two-sample t-test allowed for testing means across two groups (Global Health, 2022). Below is the portion of code used to isolate the two variables of interest, convert the character field to a factor, and perform the t-test.

B2, Output and Results

After isolating the two variables for testing, I checked the class of both variables to ensure they were the proper type for analysis. It was determined that the Churn field was a character type and Income was numeric, which was expected. Churn was converted to a factor, which resulted in two levels (D207, n.d). Income was left as-is.

Before

```
# View class/type of Churn & Income fields
class(churn_test$Churn)
```

```
## [1] "character"
```

```
class(churn_test$Income)
```

```
## [1] "numeric"
```

After

```
# Convert Churn field to factor and confirm change
churn_test$Churn <- as.factor(churn_test$Churn)
class(churn_test$Churn)
```

```
## [1] "factor"
```

```
levels(churn_test$Churn)
```

```
## [1] "No"  "Yes"
```

The next step was to run the t-test. The test would verify whether the incomes of customers who have discontinued service in the last month differ from those of active customers. The null hypothesis is that there is no difference among the means, and the alternative hypothesis is that there is a statistically significant difference between the means. The test revealed no significant difference between the two means, so the null hypothesis could not be rejected. The resulting p-value was 0.55, much higher than the necessary 0.05 to reject the null hypothesis (Global Health, 2022).

```
# Perform t-test to see if income levels differ from customers who discontinued service
#[In-text citation: (Global Health, 2017)]
t.test(Income ~ Churn, data = churn_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  Income by Churn
## t = -0.58803, df = 4601.6, p-value = 0.5565
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1644.1459   885.4205
## sample estimates:
##  mean in group No mean in group Yes
##          39706.40          40085.76
```

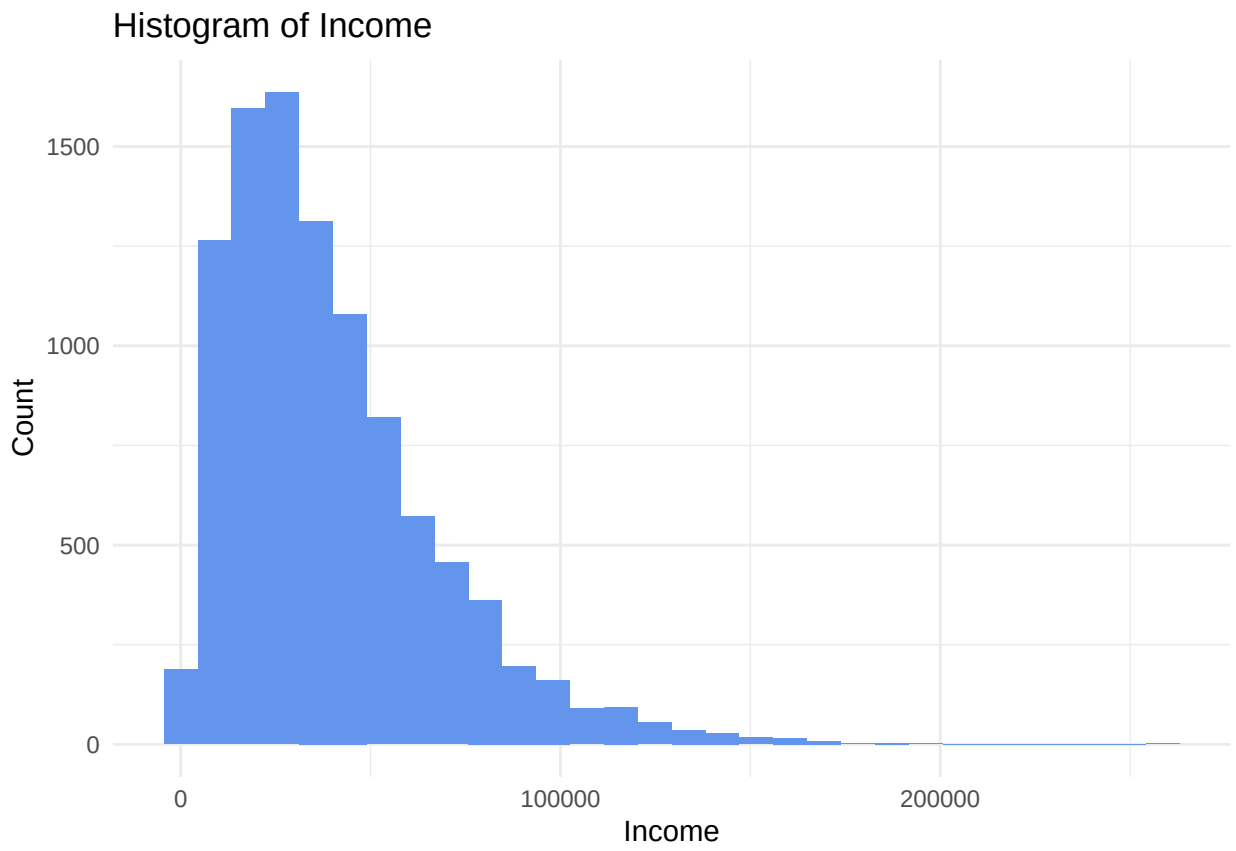
B3, Justification

Once the research question was formed, it was clear that the t-test was the best approach for testing the hypothesis. The Chi-squared technique is used for testing categorical variables, and this test revolves around the testing of a numeric variable. This approach was ruled out immediately. ANOVA tests are used when there are three or more groups to compare, which was not the case here, as Income and Churn were the only variables. Thus, the t-test was the correct technique. After choosing the t-test, I settled on a two-sided test because I was interested in determining if there was a difference in means in either direction.

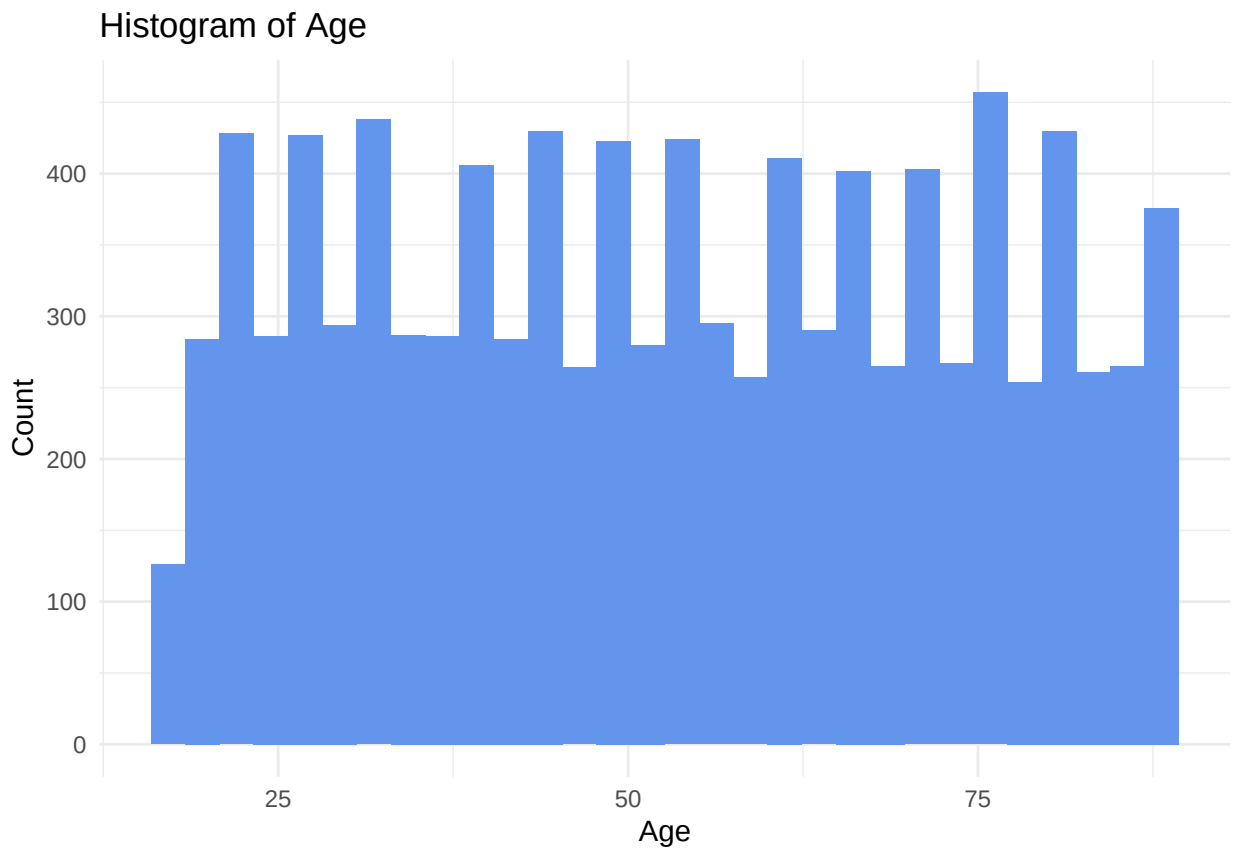
C1, Univariate Findings

Two continuous and two categorical variables were identified and analyzed using univariate statistics. Income and Age are both continuous numeric variables. Churn and Contract are both categorical variables. Histograms were created for each numeric variable to see how the values were distributed. Box plots were then created as an additional way to see the spread of values and more easily identify potential outliers. **Continuous Variables**

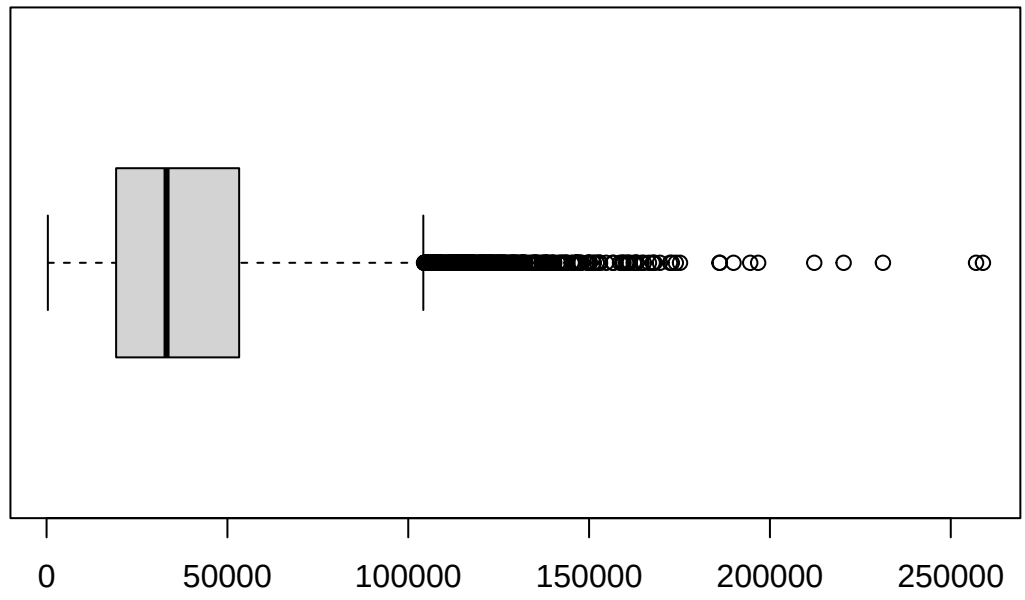
```
# Distribution of 2 continuous variables using univariate Statistics [In-text citation: (Kabacoff, n.d.
ggplot(churn_df, aes(x= Income))+
  geom_histogram(fill = "cornflowerblue")+
  labs(title = "Histogram of Income", y= "Count")+
  theme_minimal()
```



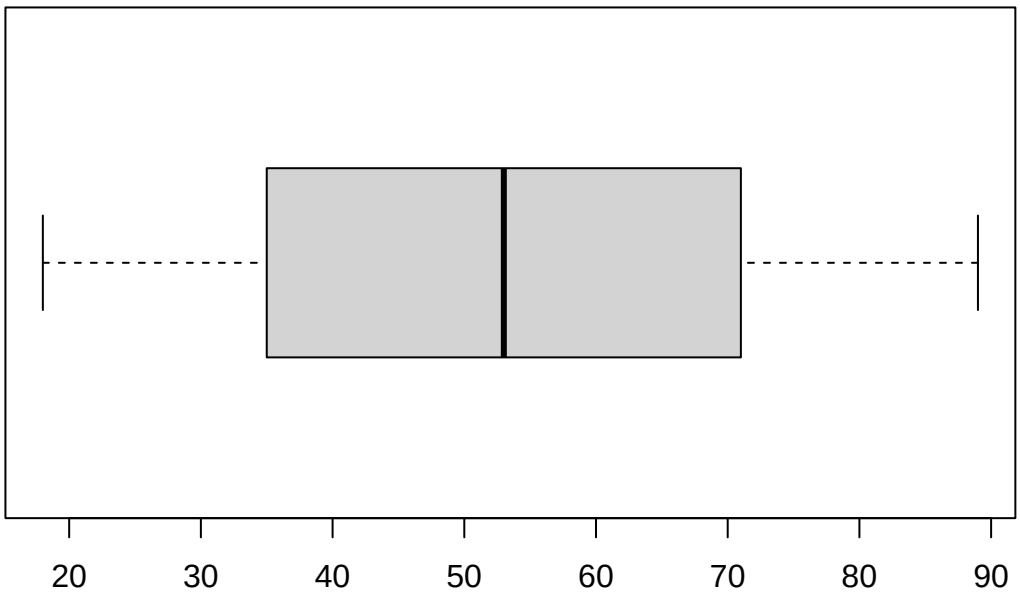

```
ggplot(churn_df, aes(x= Age))+
  geom_histogram(fill = "cornflowerblue")+
  labs(title = "Histogram of Age", y= "Count")+
  theme_minimal()
```




```
boxplot(churn_df$Income, horizontal = TRUE, main = "Distribution of Income")
```




```r
boxplot(churn_df$Age, horizontal = TRUE, main = "Distribution of Age")
```




```
stack(summary(churn_df$Income))
```

```
##      values     ind
## 1    348.67    Min.
## 2  19224.72 1st Qu.
## 3  33170.60  Median
## 4  39806.93    Mean
## 5  53246.17 3rd Qu.
## 6 258900.70    Max.
```

```
stack(summary(churn_df$Age))
```

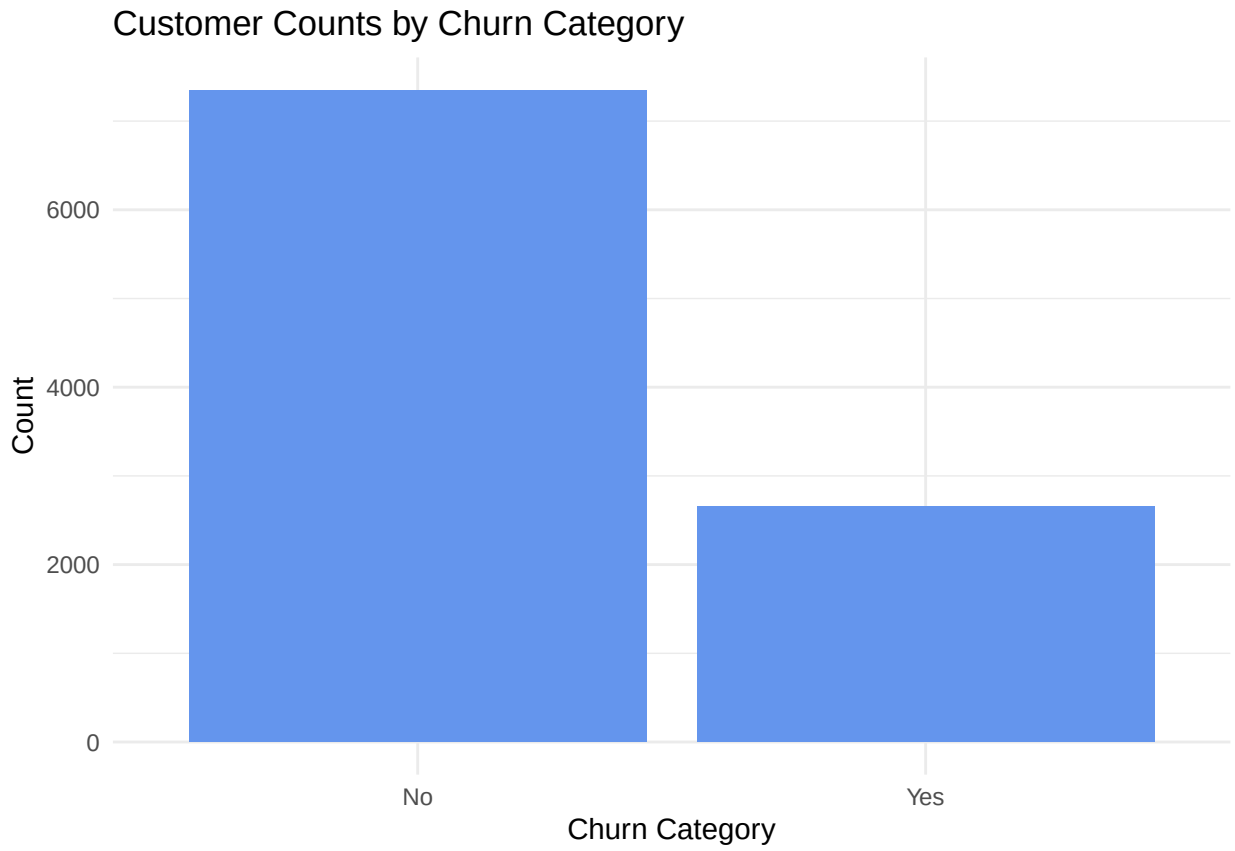
```
##    values     ind
## 1 18.0000    Min.
## 2 35.0000 1st Qu.
## 3 53.0000  Median
## 4 53.0784    Mean
## 5 71.0000 3rd Qu.
## 6 89.0000    Max.
```

One can better understand the variables by viewing the histograms and boxplots. The Income variable is right skewed, with most values clustered around the median at $33K. That can be seen in the boxplot as well, with the added benefit of seeing that there are potential outliers. The values in the upper end of the Income variable were deemed reasonable given that this spread is representative of income distributions. The distribution of the Age variable is

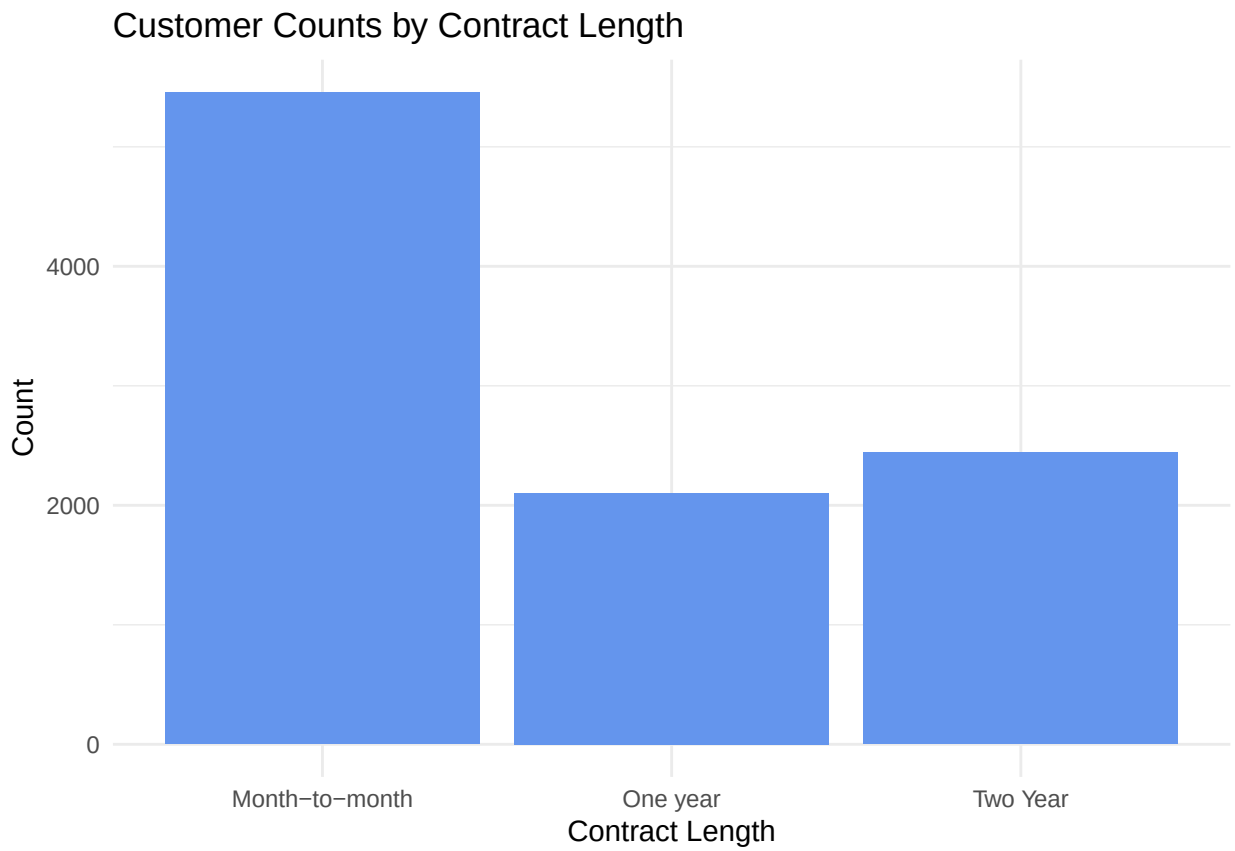
relatively uniform. The bars on the histogram are even or in the same range for the most part. The boxplot shows no outliers outside the fences, and the median is directly in the center of the box.

Categorgical Variables To visualize the two categorical variables, Churn and Contract, bar charts were utilized to view the count and distribution of the respective categories. A summary table was also produced to get the percentage of each category versus the whole. Given the nature of the Churn variable, it would be expected to have most of the values in the "No" category. The company should have more active customers than discontinued customers within the last month. The resulting chart and table for the Contract variable are compelling. Over half of the customers have a contract status of month to month, with contract lengths of one year and two years hovering around the same amount. Another potential research question to explore could be: Does contract length impact churn?

```
# Distribution of 2 categorical variables using univariate statistics [In-text citation: (Kabacoff, n.d
ggplot(churn_df, aes(x = Churn))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Customer Counts by Churn Category", x = "Churn Category",y = "Count")+
  theme_minimal()
```




```
ggplot(churn_df, aes(x = Contract), main = "Count by Contract Length")+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Customer Counts by Contract Length", x = "Contract Length",y="Count")+
  theme_minimal()
```




```
churn_df %>%
  count(Churn) %>%
  mutate(pct = n/sum(n))
```

```
##   Churn    n   pct
## 1    No 7350 0.735
## 2   Yes 2650 0.265
```

```
churn_df %>%
  count(Contract) %>%
  mutate(pct = n/sum(n))
```

```
##         Contract    n    pct
## 1 Month-to-month 5456 0.5456
## 2       One year 2102 0.2102
## 3       Two Year 2442 0.2442
```

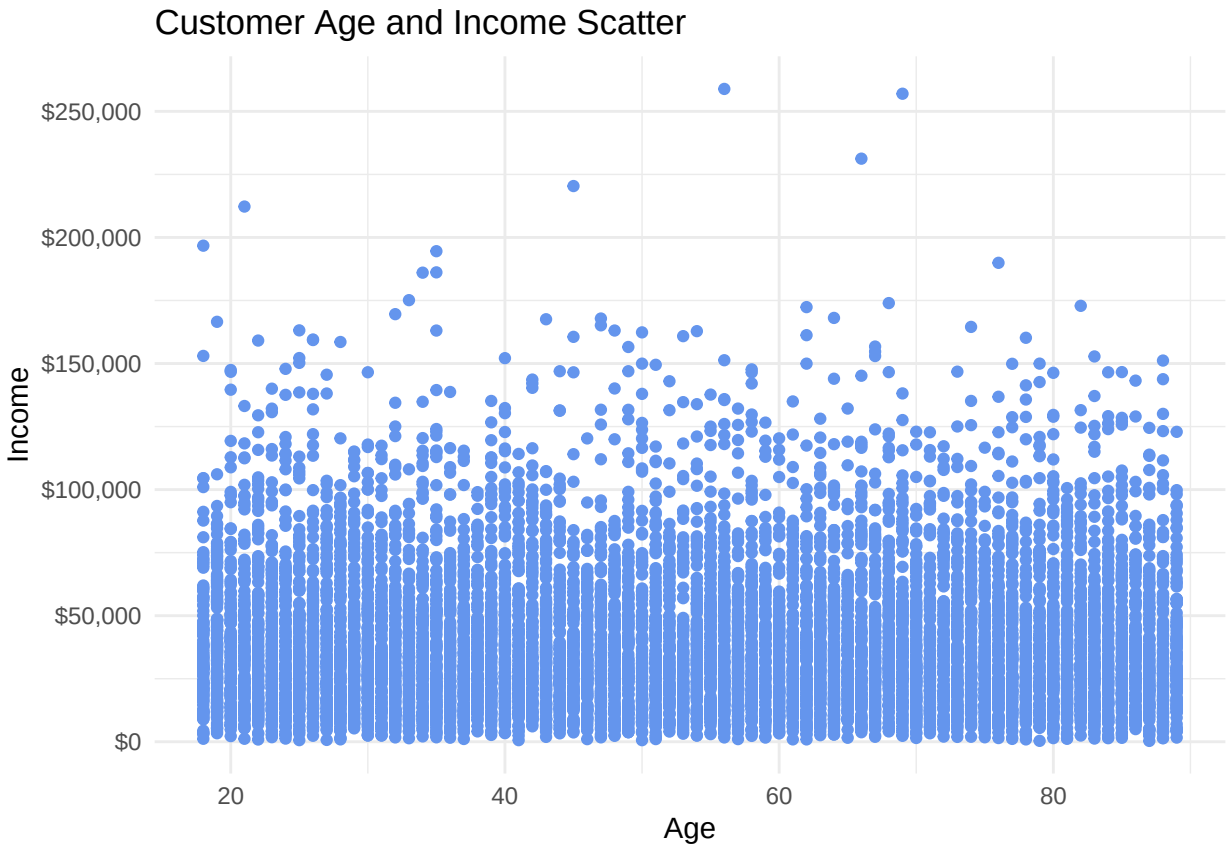
D1, Bivariate Findings

Two continuous and two categorical variables were identified and analyzed using bivariate statistics. The two continuous numeric variables chosen were Income and Age. The best way to present the combination of these variables is through a scatter plot, which would show if there were a general linear relationship between the two variables.

Summary statistic tables were also produced to give additional context to the plot. The two categorical variables chosen for analysis were Gender and Churn. A stacked bar chart was used to look at the breakout of Gender by Churn Category. A summary table was used to show the count of churn status grouped by gender and the percentage in each gender group.

Given that the distribution of the Age variable was uniform, the dots are spaced out evenly on the x-axis of the scatterplot. There is not much of a correlation between Age and Income, as most of the dots are clustered around the median income value of $33K. The stacked bar chart and ensuing summary table show that a higher proportion of male customers have discontinued service in the last month compared to those who identify as female or non-binary. Another statistical test could be run to see if the difference is statistically significant. **Bivariate Continuous**

```
# Distribution of 2 continuous variables using bivariate statistics [In-text citation: (Kabacoff, n.d.)]
ggplot(churn_df, aes(x = Age, y = Income))+
  geom_point(color = "cornflowerblue")+
  scale_y_continuous(breaks = seq(0,300000,50000),label = scales::dollar)+
  labs(title = "Customer Age and Income Scatter")+
  theme_minimal()
```




```
# Correlation Coefficient and Linear Regression
cor(churn_df$Age, churn_df$Income)
```

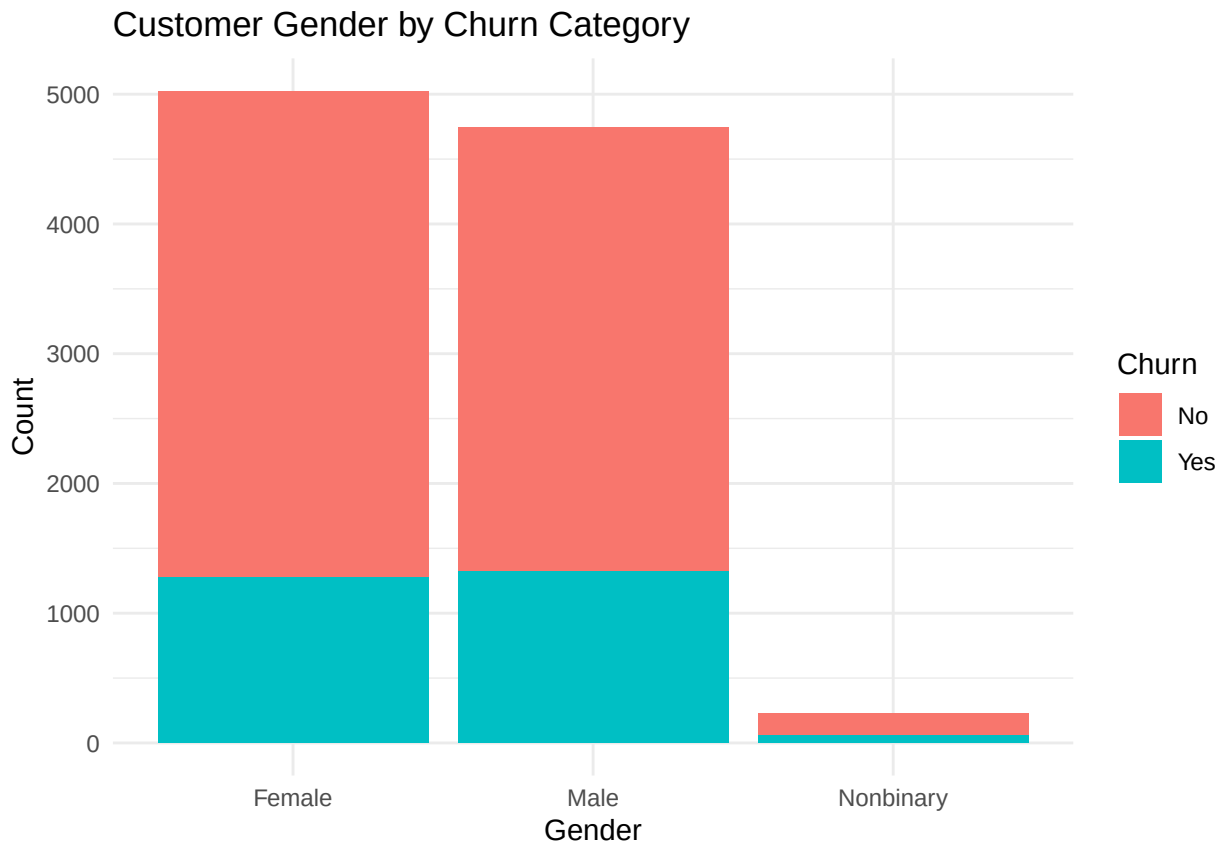
```
## [1] -0.004090602
```

```
tidy(lm(data = churn_df, Income ~ Age))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 40103.       776.     51.7     0
## 2 Age            -5.57      13.6    -0.409   0.683
```

Bivariate Categorical

```
# Distribution of 2 categorical variables using bivariate statistics [In-text citation: (Kabacoff, n.d.
ggplot(churn_df, aes(x = Gender, fill = Churn))+
  geom_bar(position = "stack")+
  labs(title = "Customer Gender by Churn Category", y = "Count")+
  theme_minimal()
```




```
# Chi-Squared Test and Contingency table
chisq.test(churn_df$Gender, churn_df$Churn)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_df$Gender and churn_df$Churn
## X-squared = 7.8801, df = 2, p-value = 0.01945
```

```
addmargins(table(churn_df$Gender,churn_df$Churn))
```

```
##
##                No   Yes   Sum
##   Female     3753  1272  5025
##   Male       3425  1319  4744
##   Nonbinary   172    59   231
##   Sum        7350  2650 10000
```

E1, Hypothesis Test

The research objective was to check whether the incomes of customers who have discontinued service in the last month differed from those of active customers. The null hypothesis is that there is no difference among the mean incomes, and the alternative hypothesis is that there is a statistically significant difference between the mean incomes. The resulting p-value was 0.55 using a 95% confidence level. In order to reject the null hypothesis and determine whether there was a statistically significant difference between the means of the groups, a p-value less than 0.05 was required. Therefore, it was determined that there was no significant difference in the incomes of customers who have discontinued service versus active customers (Global Health, 2022).

E2, Limitations

The analysis contains a large sample size of both churn categories, so the fact that there was no meaningful difference between the mean incomes indicates that other avenues should be explored to gain insights into customer churn. Because of steps taken during data cleaning, the income levels may not give an accurate representation. For example, if there were many missing or extreme values in the original dataset, an analyst may have imputed these with the median or other imputation technique. This could result in artificially low income numbers.

E3, Course of Action

Given that this was a cleaned and prepared dataset and there was not enough evidence to reject the null hypothesis, it can be concluded that the company should not focus on income levels as a predictor of churn. Additional questions should be explored to see if other variables can predict customer churn and assist with marketing. One such example may be whether the type of contract a customer has impacts churn. Over half the customers in the dataset were on month-to-month contracts, which may not inspire loyalty. The customer satisfaction survey results should also be explored. Customers who feel they are not treated fairly or have issues not resolved promptly may want to cancel services.

F, Panopto Recording

I created a Panopto video recording that covered the execution of the code and tools used in the analysis. The video link is included in the submission.

G, Sources for Code

Bobbitt, Z. (July 30, 2021). How to turn off scientific notation in R (with examples). Statology. Retrieved September 6, 2024, from (https://www.statology.org/turn-off-scientific-notation-in-r/)

Global Health with Greg Martin. (2017, June 8). R programming for beginners – statistic with R (t-test and linear regression) and dplyr and ggplot [video]. YouTube. Retrieved October 20, 2024, from (https://www.youtube.com/watch?v=ANMuuq502rE)

Kabacoff, Robert. (n.d.). Bivariate Graphs. Modern Data Visualization with R. Retrieved October 19, 2024, from (https://rkabacoff.github.io/datavis/)

Tierney, N. (n.d.). Dealing with Missing Data in R [MOOC]. DataCamp. (https://app.datacamp.com/learn/courses/dealing-with-missing-data-in-r)

H, Sources for Content

Global Health with Greg Martin. (2022, April 5). T-test, ANOVA and Chi Squared test made easy [video]. YouTube. Retrieved October 20, 2024, from (https://www.youtube.com/watch?v=ijeEYFnS2v4)

WGU College of Information Technology (n.d.). D207 Exploratory Data Analysis Webinar [Panopto Video]. Western Governors University. (https://westerngovernorsuniversity-my.sharepoint.com/:p:/g/personal/william_sewell_wgu_edu/EeImjh-L0bZImfKaMw_qluABjnC6-KJDcd-YTxCY1iIK6Q?e=A16rVL)