Who is Leaving Thanksgiving Dinner to Go Black Friday Shopping?

Date: 04-28-2021

FruityPebbles 11L - Mahish Kewalramani, Catherine Liu, Navya Belavadi

```
library(tidyverse)
library(broom)
library(performance)
library(see)
library(scales)
```

Introduction

Our data set, thanksgiving-2015-poll-data.csv, is the data behind the story "Here's What Your Part of America Eats On Thanksgiving," (https://fivethirtyeight.com/features/heres-what-your-part-of-america-eats-on-thanksgiving/) which was gathered using a Survey Monkey poll from Nov 17, 2015. 1,058 respondents answered questions about their thanksgiving habits and traditions. We were especially interested in the variable in which the respondent reported whether they would go Black Friday shopping on Thanksgiving day or not. We decided to explore the other factors that could contribute to this variable, wondering if it could reveal anything about American values of consumerism over family time or what goes into deciding whether an individual will leave Thanksgiving dinner for sales.

Research Question: Is there a relationship between household income, gender, age, and categorical region with the likelihood that an individual will go Black Friday shopping on Thanksgiving (shop early)?

Hypotheses: As household income increases to a certain extent, an individual will be more likely to go Black Friday shopping on Thanksgiving, with the exception being that extremely wealthy are less likely to go. As age increases, early Black Friday shoppers should decrease. We predict that women will be more likely to shop early. And, as we observe from rural to suburban to urban, we predict that the likelihood of a resident going Black Friday shopping will increase.

For further research, we looked to these articles: https://money.cnn.com/2015/11/30/news/companies/cyber-monday-sales/ https://news.gallup.com/poll/158927/black-friday-shopping-mostly-young.aspx

Relevant Variables:

income: the household income range, we binned into 4 categories (0 to 24,999, 25,000 to 74,999, 75,000 to 99,999, and 100,000+)

gender: male or female

age: age in years, responses were binned into 4 categories (18-29, 30-44, 45-59, 60+)

location_type: describes the region as Urban, Rural, or Suburban

black_friday: 1 if the individual responded "yes" to the question "Will you shop any Black Friday sales on Thanksgiving day," and 0 if the individual responded "no." This is the variable that we are testing to see how it is impacted by the above 4 variables.

Data manipulation: Before working with our data, we manipulated it by renaming variables to more concise names, changing blank responses to N/A to filter them out, and changing yes and no in the black_friday

variable to 1 and 0 so it would be easier to do the regression model.

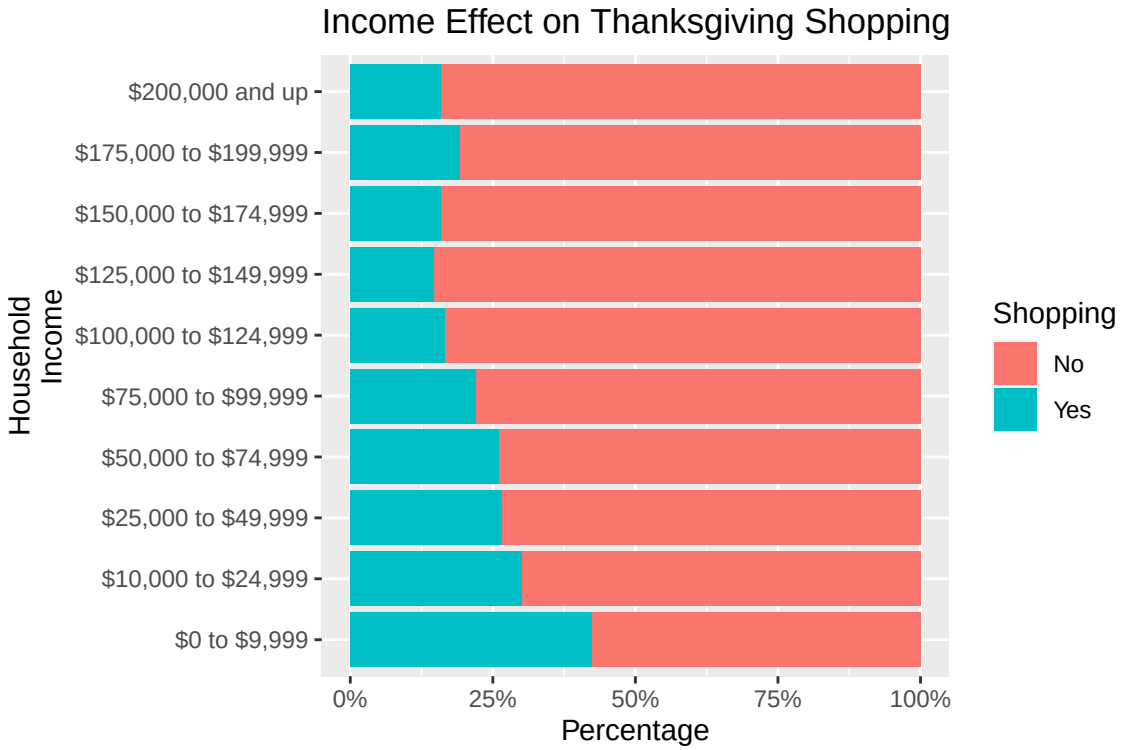
Methodology

Statistical Method: We decided to use logistic regression because our response of interest, whether an individual will go Black Friday shopping on Thanksgiving, is categorical and binary, with the responses being either "yes" or "no." We will use this logistic regression model to obtain predicted probabilities of success for our binary response variable, going Black Friday shopping early. To find the our logistic regression model with the best fit, we will look at the p-values for different variables and eliminate ones that are not significant. By doing so, we will end with a logistic regression model that ideally is the most suitable for predicting Black Friday shopping.

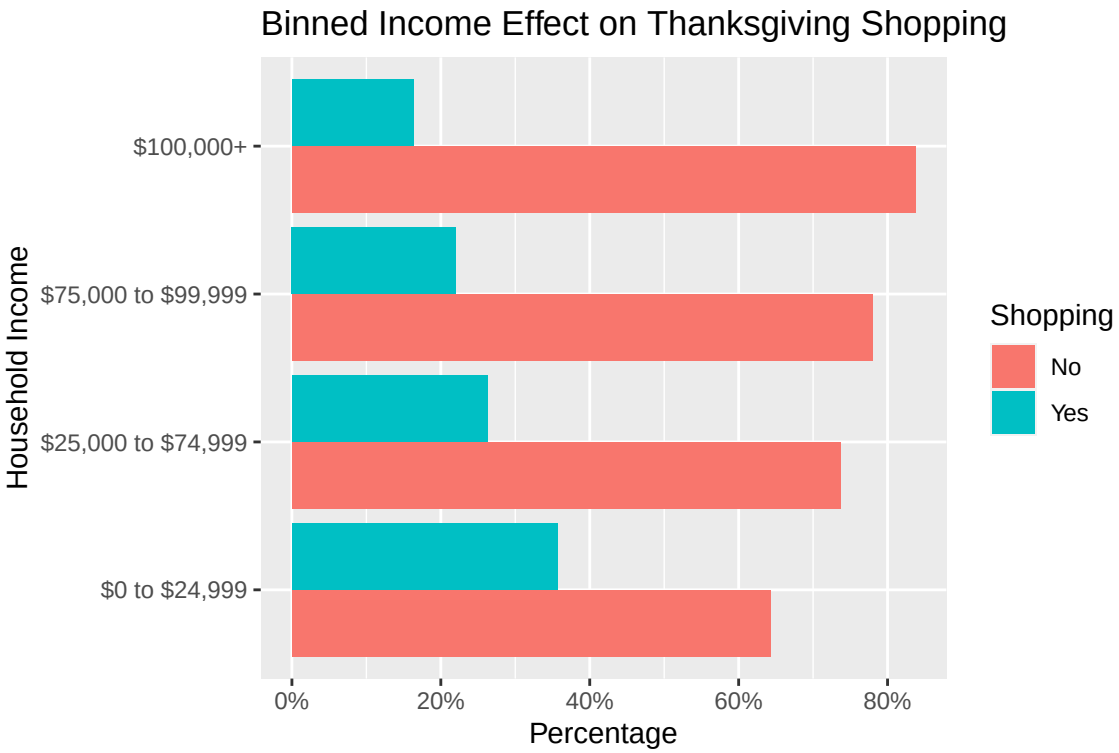
Exploratory Data Analysis

Visualization Method: For these first few graphs, we decided to use grouped and percent stacked bar charts (varying between the two based on how many groups were included for easier readability). After experimenting with other graphs, we felt this was the best way for a reader to quickly spot the relationships without unnecessary complications.

Income vs. Black Friday Shopping: First we are going to observe the different percentages within the income groups that the data provides of people that either will or will not go Black Friday shopping early.

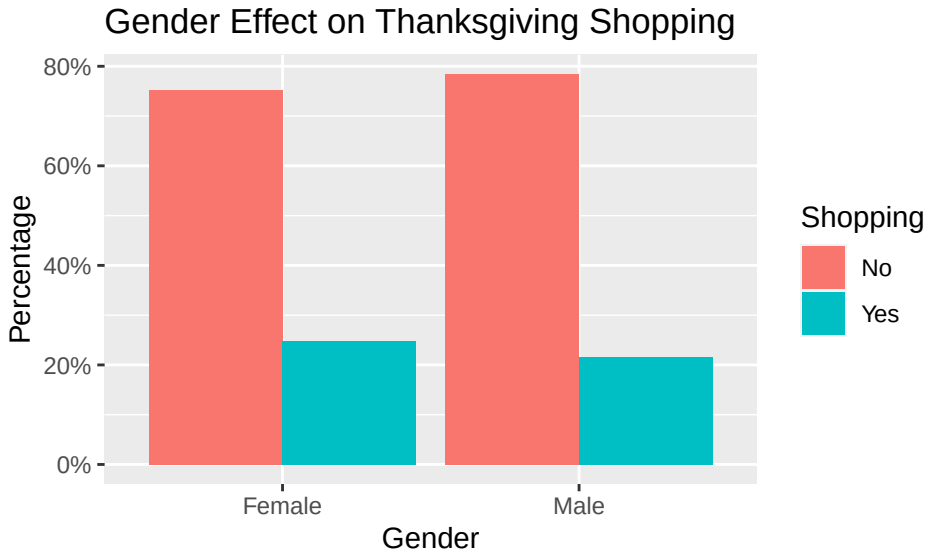


From this initial visualization of the relationship between Household Income and whether or not an individual will go Black Friday shopping early, we decided to bin the income variable further into less categories. We chose the bins of a household income from \$0 to \$24,999, \$25,000 to \$74,999, \$75,000 to \$99,999, and \$100,000+ because it appeared that the percentages naturally divide across those intervals.



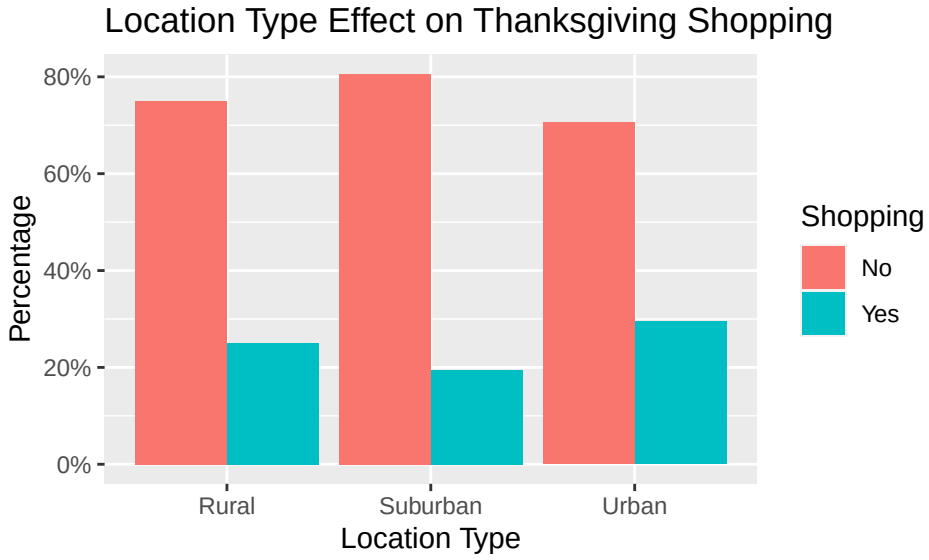
From our visualization, it appears that there is a clear relationship between household income and probability of going Black Friday shopping early – as household income increases, the percentage of people that will go Black Friday shopping decreases. The highest percentage of individuals that went shopping, at around 35%, belongs to those in the $0 to $24,999 income bracket, whereas the lowest percentage of individuals that went shopping, at roughly 18%, refers to the wealthiest income bracket ($100,000+).

Gender vs. Black Friday Shopping: Next we will visualize the percentages within gender to see whether gender affects the likelihood. We ask the question, will a greater percentage of males or females go Black Friday shopping on Thanksgiving?



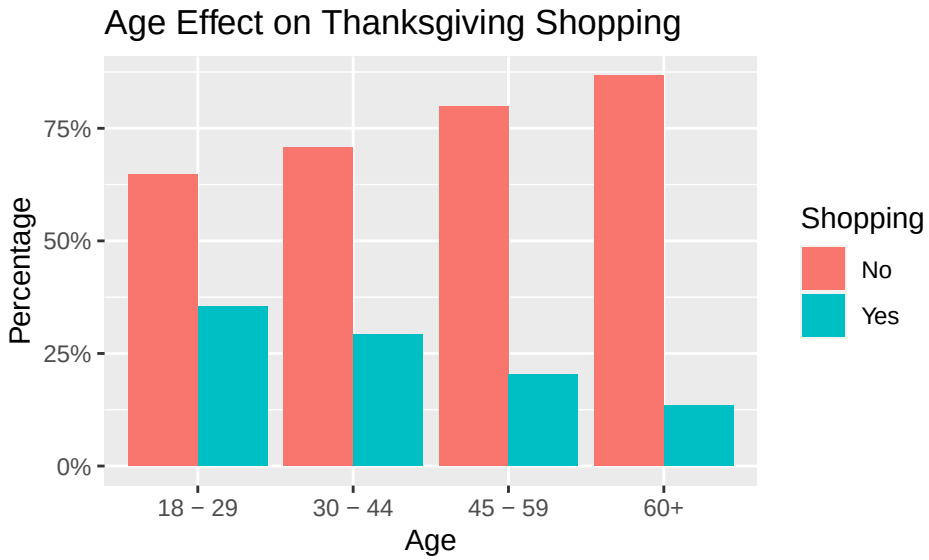
It appears that from this plot, the percentage within males and females that will or will not go Black Friday shopping early is pretty similar, with female likelihood of shopping only being slightly higher than that of males (less than 5% greater).

Location Type vs. Black Friday Shopping: Next will visualize the effect of the type of location on Black Friday shopping. The three location types we have from the data set are rural, suburban, and urban.



From our visualization, the percentage of individuals who will go Black Friday shopping across different location types does not seem to vary significantly. However, the individuals with the greatest percentage saying yes to going shopping on Thanksgiving live in Urban areas, followed by Rural and Suburban.

Age vs. Black Friday Shopping: Finally, we will visualize the percentage of Black Friday shoppers across different age groups of the survey respondents (18-29, 30-44, 45-59, and 60+).



From this visualization, we can observe that within older age groups, there is a lower percentage of people who report that they will go Black Friday shopping early. The age group 18 to 29 years old has the highest percentage of individuals who will go Black Friday shopping, while the group 60+ has the lowest. Overall, there is a clear relationship that as age increases, the likelihood of an individual Black Friday shopping early decreases.

Logistic Regression Model

We will start out with a logistic regression model that includes all the potentially interesting variables that could contribute to Black Friday Shopping early. For our first logistic regression model, we assigned Black Friday shopping as the response variable and income, gender, location type, and age as the explanatory variables.

```
# A tibble: 10 x 5
   term                      estimate std.error statistic  p.value
   <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                 -0.220     0.294    -0.748 0.454
 2 income$25,000 to $74,999    -0.188     0.251    -0.749 0.454
 3 income$75,000 to $99,999    -0.395     0.305    -1.29  0.196
 4 income$100,000+             -0.612     0.292    -2.10  0.0361
 5 genderMale                  -0.181     0.178    -1.02  0.307
 6 location_typeSuburban       -0.289     0.225    -1.28  0.199
 7 location_typeUrban           0.174     0.242     0.718 0.473
 8 age30 - 44                  -0.194     0.243    -0.798 0.425
 9 age45 - 59                  -0.600     0.261    -2.30  0.0213
10 age60+                      -1.06      0.277    -3.83  0.000127
```

Elimination of Variables:

Here we are removing gender as a variable, because it has a really high p-value and therefore seems not to have a very large effect on Black Friday Shopping. For this next logistic regression model, we have income, location type, and age as the explanatory variables.

```
# A tibble: 9 x 5
  term                     estimate std.error statistic  p.value
  <chr>                       <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                -0.300     0.283    -1.06  0.289
2 income$25,000 to $74,999   -0.184     0.251    -0.733 0.464
3 income$75,000 to $99,999   -0.405     0.305    -1.33  0.184
4 income$100,000+            -0.626     0.292    -2.14  0.0320
5 location_typeSuburban      -0.287     0.225    -1.28  0.202
6 location_typeUrban          0.163     0.242     0.674 0.501
7 age30 - 44                 -0.189     0.243    -0.780 0.436
8 age45 - 59                 -0.598     0.261    -2.30  0.0217
9 age60+                     -1.05      0.277    -3.81  0.000141
```

Next, we found that since location type has the next highest p-value, we should try eliminating that as well. For this final logistic regression model, we have income and age as the only explanatory variables since those had the lowest p values and therefore greatest impact on shopping.

```
# A tibble: 7 x 5
  term                     estimate std.error statistic  p.value
  <chr>                       <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)                -0.342     0.225    -1.52  0.128
2 income$25,000 to $74,999   -0.234     0.249    -0.941 0.347
3 income$75,000 to $99,999   -0.472     0.303    -1.56  0.119
4 income$100,000+            -0.719     0.288    -2.50  0.0126
5 age30 - 44                 -0.161     0.242    -0.668 0.504
6 age45 - 59                 -0.606     0.260    -2.34  0.0195
7 age60+                     -1.05      0.276    -3.82  0.000134
```

Model Performance:

We decided to analyze the performance of our 3 logistic models by looking at the Bayesian Information

Criterion(BIC) because it takes into account both model performance and model complexity. For comparing models, the smaller the BIC score, the better the fit is.

Logistic Regression Model 1 Performance:

```
# Indices of model performance

AIC     |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
-------------------------------------------------------------------------------------------------
808.673 | 854.914 |     0.054 | 0.416 | 1.030 |    0.524 |   -50.440 |           0.008 | 0.654
```

Logistic Regression Model 2 Performance:

```
# Indices of model performance

AIC     |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
-------------------------------------------------------------------------------------------------
807.720 | 849.337 |     0.052 | 0.416 | 1.030 |    0.524 |   -50.394 |           0.004 | 0.654
```

Logistic Regression Model 3 Performance:

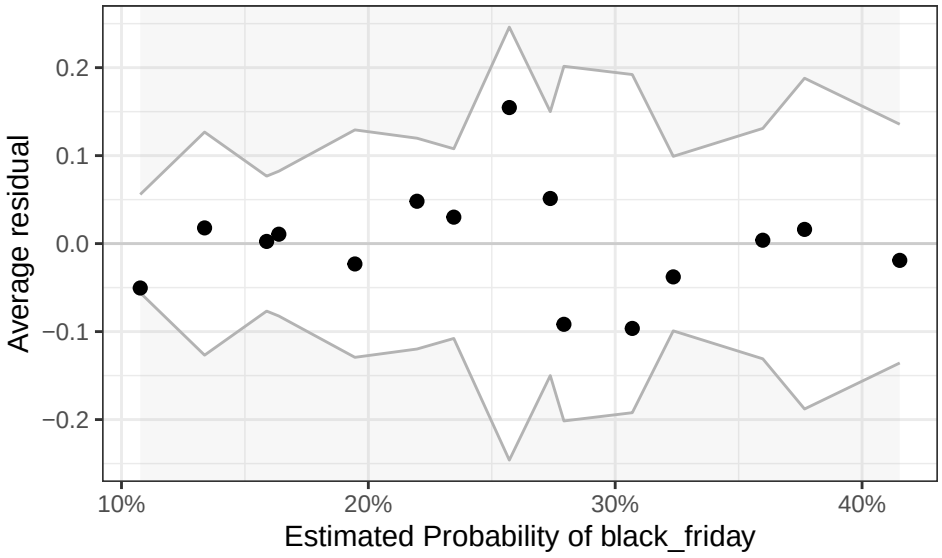
```
# Indices of model performance

AIC     |     BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
-------------------------------------------------------------------------------------------------
808.729 | 841.098 |     0.046 | 0.418 | 1.032 |    0.528 |   -50.172 |           0.005 | 0.652
```

From looking at the respective BIC for our 3 models, we decided to select our third model. The BIC statistic, 841.098, is the lowest out of the three, and therefore we can assume that it is has the best performance, while not being overly complicated. In our case, BIC proved to be the best measure of model fit to use because it actually appropriate for our logistic regression models and it adjusts for both model performance and complexity.

Binned Residuals Plot:

After deciding that our third model was the most suitable in predicting early Black Friday shoppers, we decided to make a Binned Residuals plot to assess the fit of the model further. The plot is created by dividing the data into bins based on their fitted values, and then plotting the average residual versus the average fitted value for each bin. If the model is true, we expect 95% of the residuals to fall inside the error bounds.

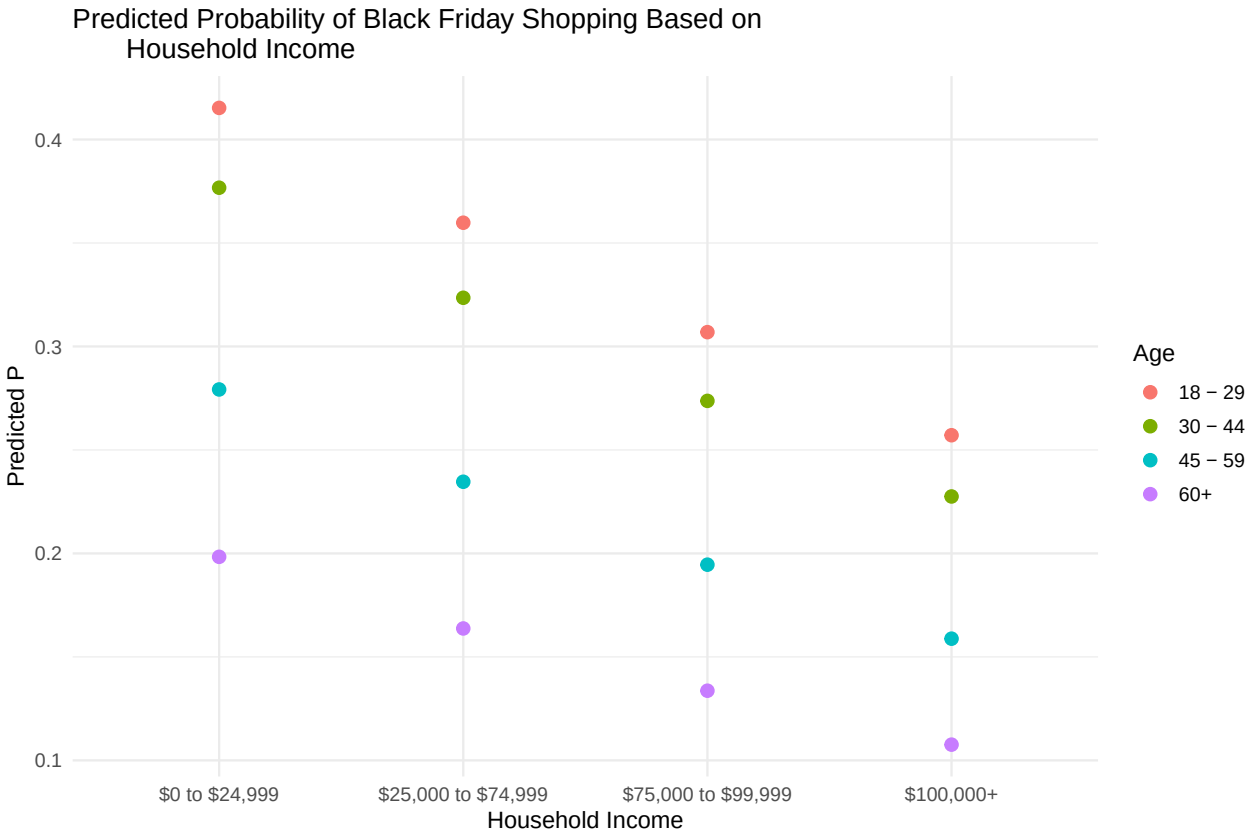
```
Ok: About 100% of the residuals are inside the error bounds.
```



Because 100% of our residuals are inside the error bounds, we can confirm that this model is a good fit.

Predicted Probabilities:

Finally, we decided to make a predicted probabilities plot to show the effect of age and income on Black Friday shopping early.



Results

After our data analysis, we concluded that there is a significant relationship between household income and age in whether an individual will go Black Friday shopping on Thanksgiving. While there was a slightly greater percentage of females than males going shopping early and the greatest percentage of individuals shopping early were from an Urban area (by a very small margin), gender and location type had little to no impact in comparison to household income and age. We found that the predicted probability of going Black Friday shopping early decreases with a higher household income and age.

Main results From our analysis we obtained the logistic regression model:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.342 - 0.234\ (income25,000to74,999) - 0.472\ (income75,000to99,999) - 0.719\ (income100,000+) -$$

$$0.161\ (age30-44) - 0.606\ (age45-59) - 1.054\ (age60+)$$

We can configure the model to get a model for $\hat{p}$ is the probability that some event occurs with our event being that an individual will shop.

$$\hat{p} = \frac{e^{-0.342-0.234\ (income25{,}000to74{,}999)-0.472\ (income75{,}000to99{,}999)-0.719\ (income100{,}000+)-}}{1+e^{-0.342-0.234\ (income25{,}000to74{,}999)-0.472\ (income75{,}000to99{,}999)-0.719\ (income100{,}000+)-}} \frac{0.161\ (age30-44)-0.606\ (age45-59)-1.054\ (age60+)}{0.161\ (age30-44)-0.606\ (age45-59)-1.054\ (age60+)}$$

We can also put our interpretation in odds to determine $\frac{\hat{p}}{1-\hat{p}}$,the odds that the event occurs.

$$\frac{\hat{p}}{1-\hat{p}} = e^{-0.342-0.234\ (income25{,}000to74{,}999)-0.472\ (income75{,}000to99{,}999)-0.719\ (income100{,}000+)-}$$
$$0.161\ (age30-44)-0.606\ (age45-59)-1.054\ (age60+)$$

Our intercept of -0.342 can be interpreted as the log-odds of Black Friday shopping for an individual who is 18 to 29 years old and in the \$0 to 24,999 income bracket. The odds of Black Friday shopping is $e^{-0.342} = 0.710$ We can use our predicted probability plot to estimate the probability of certain individuals going shopping. For example, the predicted probability of an individual between 45-59 years old with a household income of \$100,000+ shopping early is 0.159, so about 15.9 %. The predicted probability of going Black Friday shopping early for an individual aged 15-29 years old with a household income of \$0 to \$24,999 is 0.415, so about 41.5 %.

Coefficients Interpretation:

- Holding age constant, for someone in the \$25,000 to 74,999 income bracket, we expect log-odds of Black Friday shopping to decrease by 0.234. That is, the odds of Black Friday shopping is expected to be multiplied by
  $e^{-0.234} = 0.791$
- Similarly holding age constant, for someone in the \$75,000 to 99,999 income bracket, we expect log-odds of Black Friday shopping to decrease by 0.472. That is, the odds of Black Friday shopping is expected to be multiplied by $e^{-0.472} = 0.6237$.
- And for someone in the \$100,000 income bracket, we expect log-odds of Black Friday shopping to decrease by 0.719 and odds of Black Friday shopping is expected to be multiplied by $e^{-0.719} = 0.487$.
- Holding Income constant for someone in the 30-44 year old age bracket, we expect log-odds of Black Friday shopping to decrease by 0.161. That is, the odds of Black Friday shopping is expected to be multiplied by $e^{-0.161} = 0.851$.
- Holding Income constant for someone in the 45-59 year old age bracket, we expect log-odds of Black Friday shopping to decrease by 0.606. That is, the odds of Black Friday shopping is expected to be multiplied by $e^{-0.546} = 0.546$.
- Finally, holding Income constant for someone in the 60+ year old age bracket, we expect log-odds of Black Friday shopping to decrease by 1.054. That is, the odds of Black Friday shopping is expected to be multiplied by $e^{-1.054} = 0.349$.

How we arrived at our answer We arrived at our best fit model by eliminating variables based on whether they had a high p-value or not, and then analyzing the BIC scores for the three logistic regression models we made. We found that our final model had the lowest BIC value, and concluded that it had the best model performance while not being overly complex. To perform this backwards elimination, we started out including each relevant variable in our logistic model that we thought offered an interesting statistical analysis. From this technique, we decided to first eliminate gender, as the coefficient gender had a high p-value of 0.307. This is also shown in our initial exploratory data analysis, as you can see that gender did not have that much of an effect on whether an individual went Black Friday shopping early. That is, both males and females had about the same percentage of the sex that said they would go Black Friday shopping. After eliminating gender, we created a new logistic regression model. From this next model, we decided to eliminate location_type, as it had the next highest p-values. Through eliminating the variables that are not statistically significant,

we ideally found the most suitable model to predict the odds of an individual going Black Friday shopping early. After getting this final model, we analyzed it further by looking at the BIC scores, finding that our last model had the lowest one, and the binned residuals plot. The low BIC score compared to the other models showed that our final model was not overly complex and had the best performance. The results from our binned residuals plot, with all over our residuals falling within the error bounds, further confirmed that our model was a god fit. Through our analysis we could be confident that our logistic regression model with only income and age as variables was the best fit in predicting thanksgiving shopping.

Discussion

Discussion - Summary of what we learned about research question

From our analysis, we learned that age and income are the most relevant variables when it comes to determining whether an individual will go Black Friday shopping early. We can infer that individuals with a household income of $0 - $24,999 and individuals between the age range of 18 to 29 years old are the most likely to go Black Friday shopping on Thanksgiving. Our research question was what is the relationship between household income, age, gender, and location type with likelihood that someone will go Black Friday shopping on Thanksgiving? While our initial hypothesis was that as income increases a greater percentage of individuals will shop, our results showed that the exact opposite was the case. We learned that age and household income are the most relevant variables, as there was a relationship for age (as age increases, shopping probability decreases) and income (as income increases, shopping probability decreases) but not much of a difference between male's and female's likelihood of shopping or between individuals in urban vs rural or suburban areas.

Data explanation: As household income increases, individuals will not necessarily care as much about sales or require them to purchase things, and therefore will be less likely to rush to stores early for lower prices, so the income trend makes sense. Older generations also tend to prefer spending time with their family at Thanksgiving dinner, whereas younger generations prefer going out and shopping so the age trend makes sense. Geographically speaking, there should be more stores in urban areas than rural and suburban areas, and culturally speaking, shopping tends to be more of a hobby in urban areas than rural and suburban, so the (barely) greater likelihood of urban residents shopping early makes sense. Overall, the percentages of individuals saying yes to going Black Friday shopping on Thanksgiving was fairly low in 2015, which could be explained by the surge of shopping becoming more heavily online due to Cyber Monday sales.

Limitations: Critiquing our Methods While our use of eliminating values based on p-values and finding the model with the lowest BIC showed that there was some improvement with our respective models as we narrowed the relevant variables down to age and income, none of the changes were incredibly drastic to say that one was exceptionally better than the others. We found that comparatively, our third seemed like it was slightly better than the others.

When looking back at our methods, using logistic regression comes with assumptions of independence and linearity in the log-odds. However, our data may violate the independence assumption since we're analyzing survey data which is not the most random sample and is subject to bias from optional response. So, there could be some reason to say that the data points were correlated since it was not entirely random. There are also issues with the reliability and validity of our data, as it was gathered from a survey. The survey responses may not be fully reliable as they are subject to opinion bias. Many of the survey questions were asking about the respondents opinion (i.e How would you describe where you live?) so it was mainly just based on what they thought and could not be entirely valid or reliable. In terms of linearity, an issue may arise if the predictors are correlated, such as if location has to do with household income, which would lead to the coefficient estimates being unreliable.

There may be further shortfalls in our methodology with our usage of BIC. With BIC, the approximation is only valid for a sample size that is the number of data points larger than the number of parameters in the model and BIC cannot handle models complex in variable selection. Additionally, because BIC penalizes for complexity more heavily, the low score for our third, and least complex model, could also be attributed to it having less variables than our first and second models.

We did not look at the pseudo R squared value generated by model performance for out logistic regression models as we decided it was not a very useful or appropriate metric in our case. Additionally, because it is not adjusted R squared, this value increases with additional variables, and is not a good measure of the fit of our models.

What we would do differently: So, if we were to start over with the project and wanted to use more variables, We could extend to methods beyond BIC and looking at the p-values of variables to analyze our logistic regression model. In terms of independence, these critiques are simply based on the data set that we found so we could possibly bring in other datasets that have more info. For thinking about the further shortfalls of our model and where we could take the model further, we could also look at interaction effects and include more variables. Overall, there is evidently a lot more that goes into whether an individual would go Black Friday shopping early on Thanksgiving day or not beyond the ones we had access to in this set which could pose for interesting statistical analysis in the future.