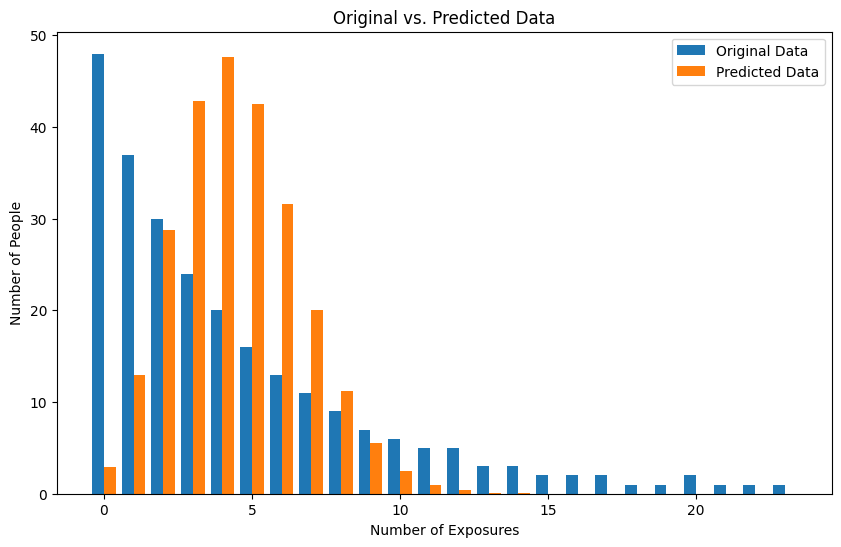
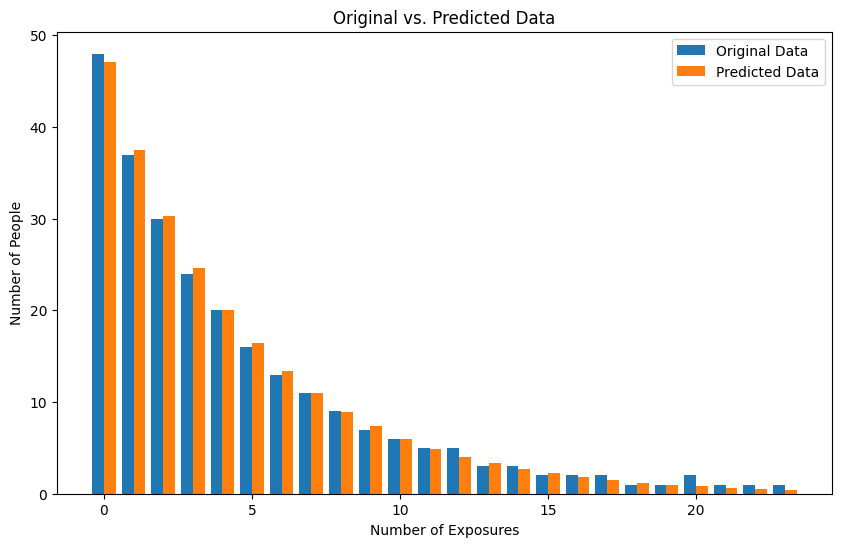
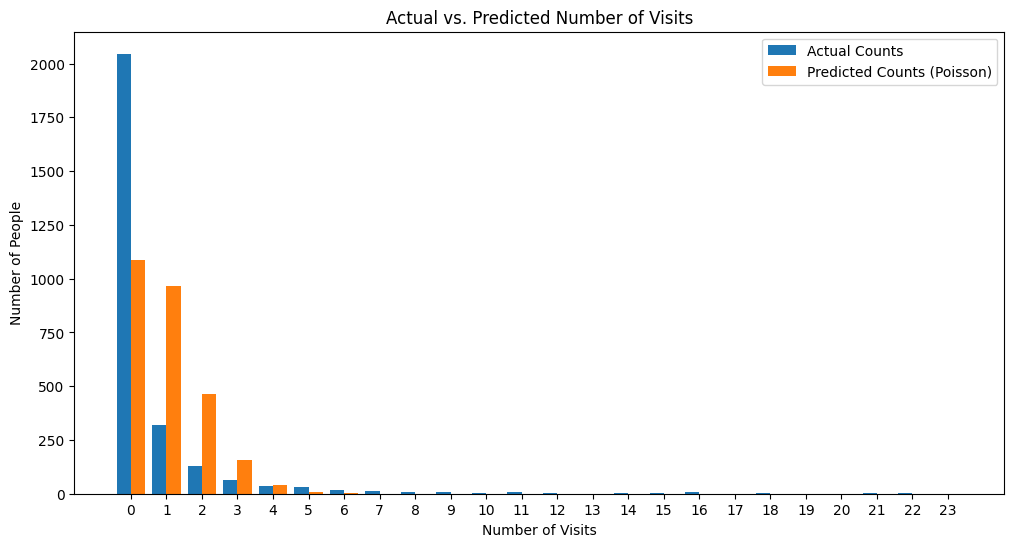
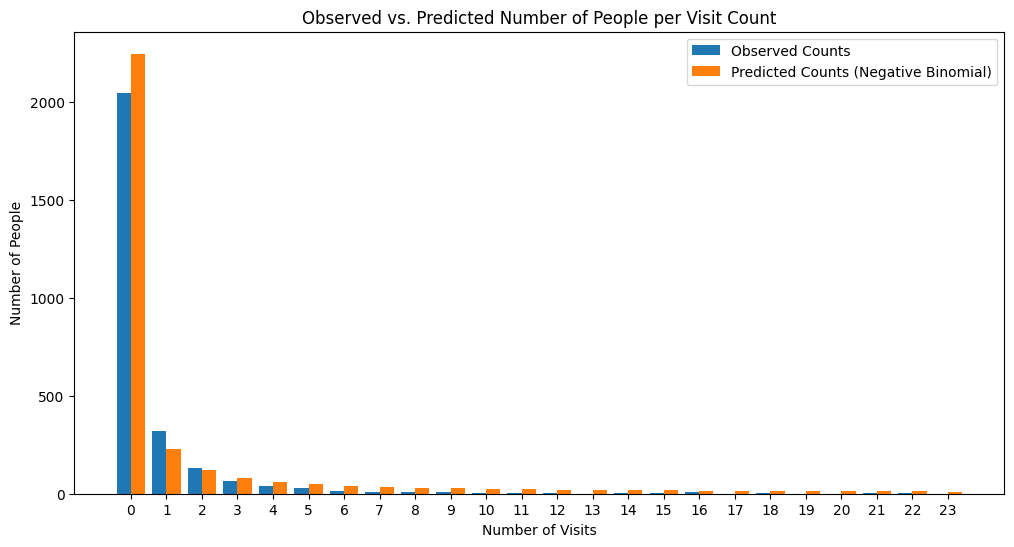
Group 1

Project 1

Names:

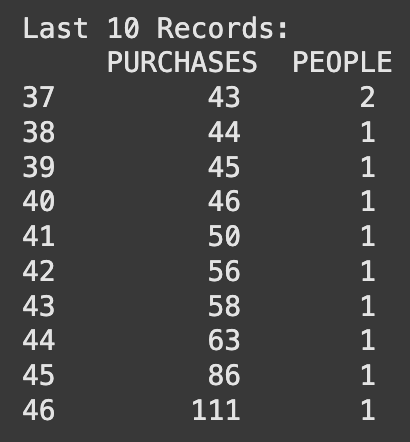
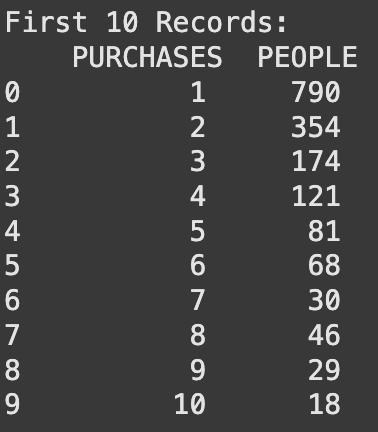
**Part I: Replicating Models From Class**

1. **The Poisson Model:** The estimated parameter, lambda, for the Poisson Model using maximum likelihood estimation was 4.456. The maximum log-likelihood was -929.044. With this estimated parameter, we were able to predict the number of people that were exposed to the billboard for exposures 0 through 23. We predicted the number of people by finding the probability of k number of exposures and then multiplied the probability by the the number of people in the dataset. For example, for 2 exposures we found the probability to be 0.115. We then multiplied this by 250 (the number of people in the dataset) and found that the predicted number of people that were exposed to the billboard twice was about 28.813 people. Below is a graph of the original data versus our predicted data:From this graph, we can see that this is not a great estimator as we are seeing very few predicted people for 0 exposures.
2. **The NBD Model:** The estimated parameters, shape and scale, for the NBD model were 0.969 and 0.179 respectively, using a maximum likelihood estimation. The maximum log-likelihood was -649.689. With these parameters we were able to predict the number of people who were exposed to the billboard in a range from 0 to 23. This was conducted by finding the probability of k exposures and multiplying it by the number of people in the dataset. For example with k = 2, we found the probability to be 0.121 and multiplied this by 250 (the number of people in the dataset) and predicted 30.319 people to be exposed to the billboard twice. Below is the graph of original versus predicted values:We can see that this is a significantly better fit to the original data than the previous Poisson model. We are seeing a larger log-likelihood with the NBD than Poisson. The NBD model is also more accurately predicting the number of people who will not be exposed to the billboard, which is a large portion of the dataset.
3. **The Poisson Regression:** The estimated parameters for the Poisson Regression were lamba0 = .0439, beta1 = 0.0938, beta2 = .0043, beta3 = 0.5883, and beta4 = -0.0359 using maximum likelihood estimation. With this the maximum log-likelihood was -6291.497. We were able to predict the number of people for website visits ranging from 0 to 23 using these parameters by calculating lambda\_i for each row with k visits and taking a sum of the probabilities for each k with the given lambda\_i of each row. For example when predicting 2 website visits, we used the Poisson probability function with k = 2 visits for all lambda\_i. Then we summed all the the results of this probability function to find the expected number of visits. Below is a graph of of predicted number of visits versus the actual number of each visit. From this graph we can see that the Poisson regression is under predicting 0 visits and over predicting 1 through 3 visits.
4. **The NBD Regression:** The estimated parameters for the NBD Regression were shape = 0.1387, scale = 8.1959, beta1 = 0.0734, beta2 = -0.0094, beta3 = 0.9022, beta4 = -0.0243, using maximum likelihood estimation. The maximum log-likelihood was -2888.966. With these parameters, we were able to predict the number of people that would visit the site ranging from 0 to 23 times. These predictions were calculated by using the NBD probability equation for k visits. This probability was then used in the NBD probability function for k visits. These probabilities were then summed to find the expected number of people visiting the site k times. Below is the graph of predicted versus actual website visits.We can see that this is a much better fit of the data than the Poisson regression. This model appears to be accounting for more 0 visits people than Poisson. We can also say the NBD is a better model due to the log-likelihood being higher than the Poisson (-2888.966 > -6921.497).
5. **Managerial Takeaways:** From these models, there is a lot of insight that we can conclude on the nature of the models ability to predict. In both cases, it seems that the Poisson models under predicted values of 0 more than their NBD counterparts. While these Poissons were not the best fit in these use cases, we can see how our data being highly right skewed with most of our data having 0 exposures/visits may have had something to do with that. For situations that the data does not include so many instances of 0, the Poisson could be a better fit.

With that being said, it appears that in both cases above that the NBD was the far superior model. They were able to predict the number of 0s in the dataset significantly better than the Poissons. However, It is hard to tell with just these two situations if an NBD would be as successful for datasets that did not see such large instances of 0. So, while the NBD performed better in these two situations, it is unfair to say that it is a better performing model overall than Poisson.

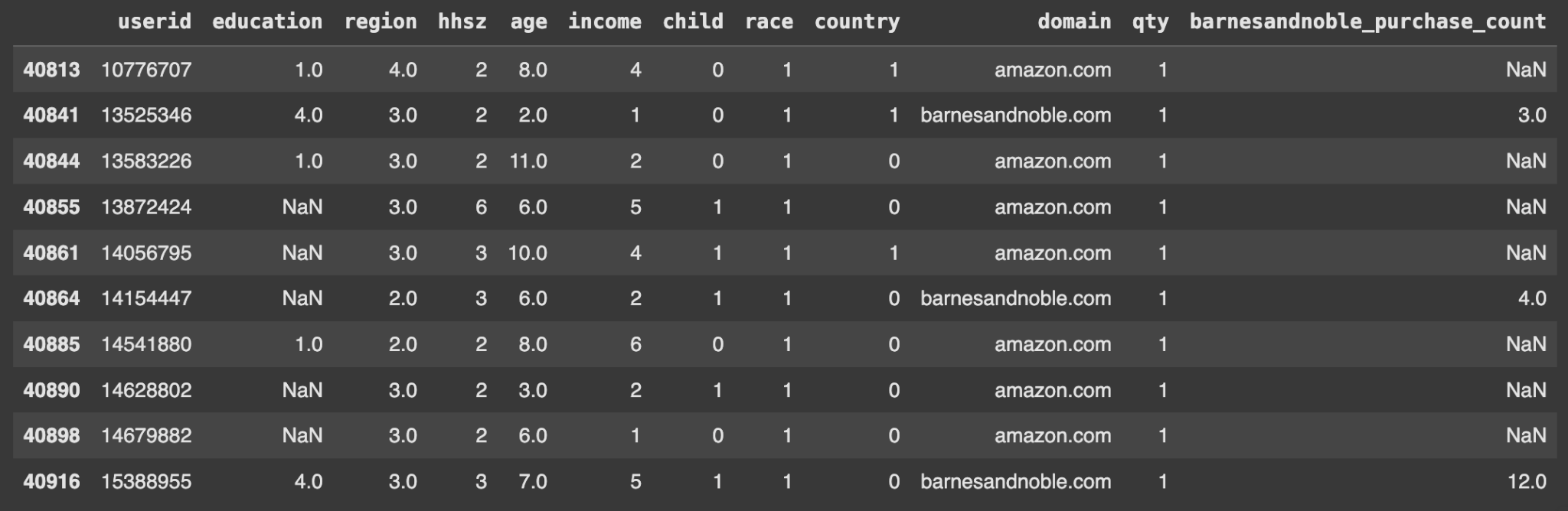
**Part II: Analysis of New Data**

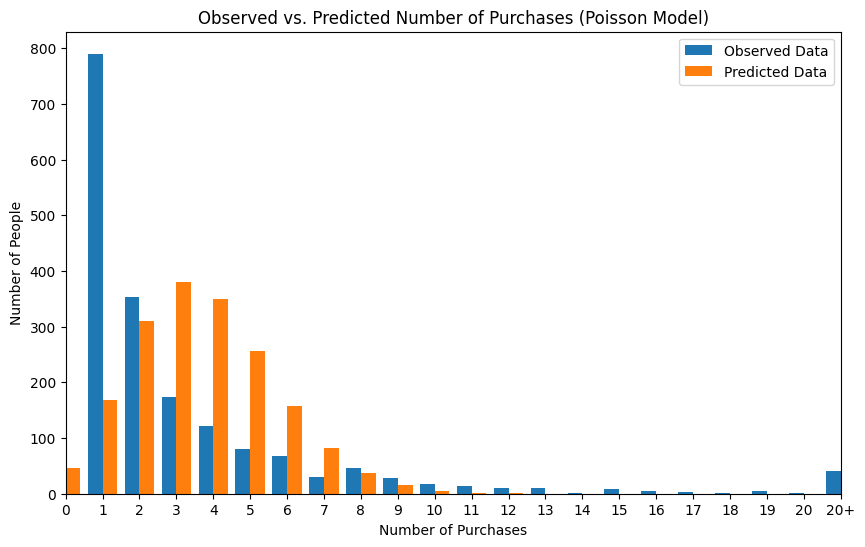
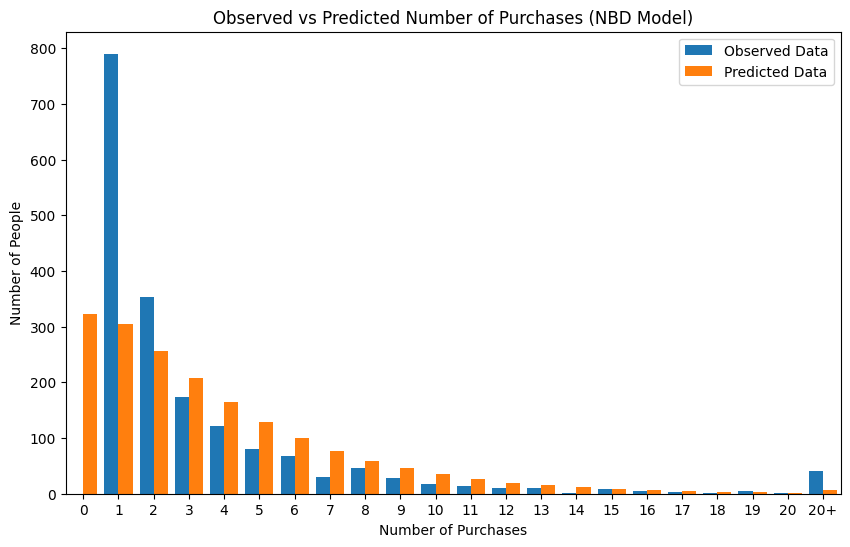
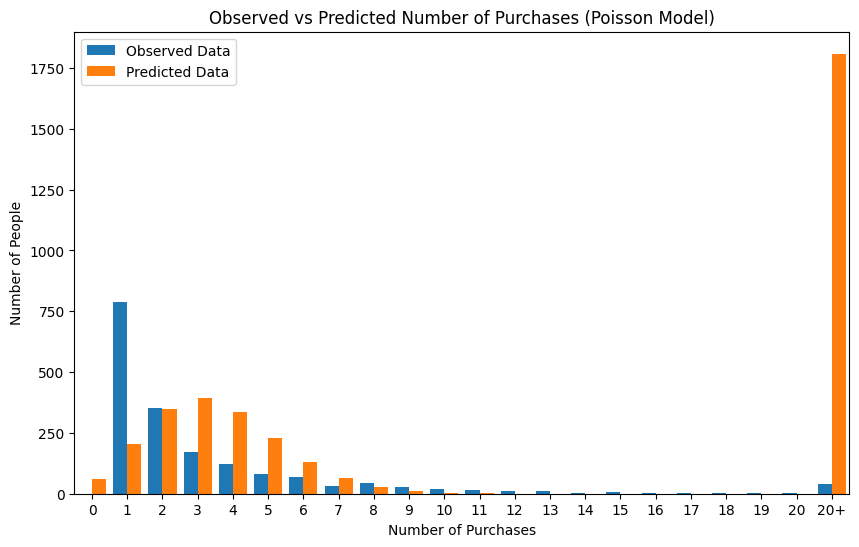
1. **Generating two new dataset:** Here are screenshots of the first and last few rows of the books01 and books02 data sets
   * Books01

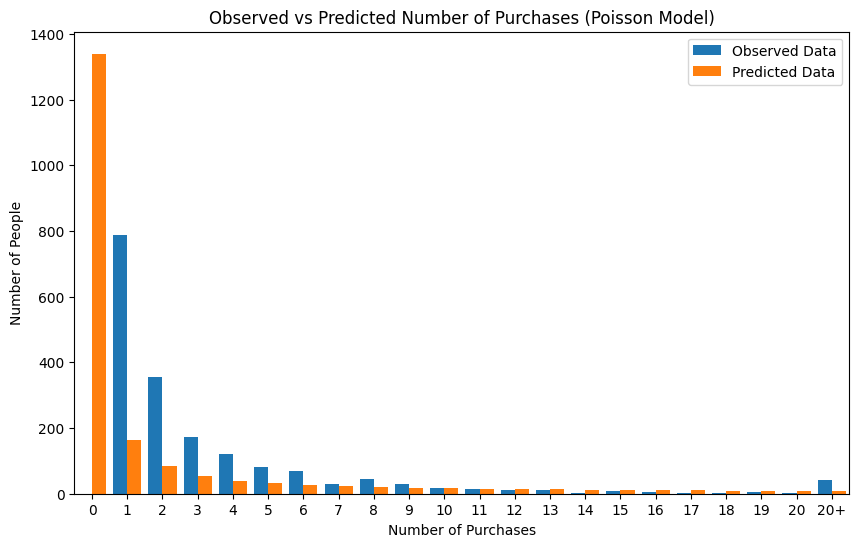


* + Books02





1. **Poisson Model on Books01:** The estimated parameter (lambda) for the Poisson model using MLE is about 3.674, while the log-likelihood is -6777.839.
2. **Poisson Model on Books02:** The estimated parameter (lambda) for the Poisson model using MLE is about 3.674, while the log-likelihood is -6777.839. As we can see we achieved the same results on Books02 as we did on Books01. Below is the graph of original versus predicted number of purchases. 
3. **NBD Model on Books01:** The estimated parameters, shape and scale for the NBD model using MLE is about 1.265 and 0.256 respectively, while the log-likelihood is -4379.262.
4. **NBD Model on Books02:** The estimated parameters, shape and scale for the NBD model using MLE is about 1.265 and 0.256 respectively, while the log-likelihood is -4379.262. Which is identical to the results that we saw on Books01. Below is a graph for predicted vs actual number of purchases. 
5. **Reach, Frequency, and GRP calculations:**
   * Reach = 1488.542
     + This was calculated by taking a sum of the predicted number of people in each purchase quantity greater than 1
   * Average Frequency = 0.0553
     + This was calculated by dividing the average number of purchases by the reach
   * Gross Rating Points = 4.545
     + This was calculated by multiplying the reach % by the average frequency
6. **Handling Missing Values:** 
   * Step 1: we identified the missing values in each dataset
   * Step 2: we defined the thresholds for “many” and “small” amounts of missing data
     + Many = more than 30%
     + Small = more than 5%
   * Step 3: we dropped variables with “many” missing values
   * Step 4: we dropped rows with “small” missing values
   * Step 5: replaced the remaining missing values with the means of their column
7. **Poisson Regression:** The estimated parameters for the Poisson Regression were lamba0 = 3.414, beta\_region = 0.0013, beta\_hhsz = .0.0214, beta\_age = 0.0168, beta\_income = 0.0144, beta\_child = -0.0424, beta\_race = -0.1110, and beta\_country = -0.2410 using maximum likelihood estimation. With this the maximum log-likelihood was -4379.262. We were able to predict the number of purchases ranging from 0 to 23 using these parameters by calculating lambda\_i for each row with k purchases and taking a sum of the probabilities for each k with the given lambda\_i of each row. For example when predicting the number of 2 purchases, we used the Poisson probability function with k = 2 purchases for all lambda\_i. Then we summed all the the results of this probability function to find the expected number of people making 2 purchases. Below is a graph of of predicted number of purchases versus the actual number of each visit. The Poisson Regression does not seem to estimate any number of purchases all that well. It also seems to be overestimating 20+ purchases and underestimating purchases less than 2. This model’s predictions appear to be most impacted by race and country as their coefficient (betas) have the highest magnitude

1. **NBD Regression:** The estimated parameters for the NBD Regression were shape = 1.2680, scale = 0.4450, beta\_region = 0.0 , beta\_hhsz =0.015, beta\_age = 0.0180, beta\_income = 0.0165, beta\_child = 0.0, beta\_race = , and beta\_country = 0.0 using maximum likelihood estimation. With this the maximum log-likelihood was -4370.209. We were able to predict the number of purchases ranging from 0 to 23. These predictions were calculated by using the NBD probability equation for k purchases. This probability was then used in the NBD probability function for k purchases. These probabilities were then summed to find the expected number of people visiting the site k times. Below is the graph of predicted versus actual website visits.Visually, we can see that this is a much better fit than the Poisson Regression as it does not overestimate the 20+ group of purchases. This model’s predictions appear to be impacted most by household size, age, and income as these coefficients (betas) have the highest magnitude.
2. **Comparison of Models**
   * Poisson Model
     + Aic = 13557.677870853298
     + BIC = 13559.528018455008
   * NBD Model
     + Aic = 8762.52321361003
     + BIC = 8776.83096528002
   * Poisson Regression
     + Aic = 13480.365565648875
     + BIC = 13524.369801532037
   * NBD Regression
     + Aic = 8758.418805263762
     + BIC = 8807.92357063232
   * Based on the AIC, the NBD Regression is the best performing model. Based on BIC, the NBD Model is the best performing. Next lets look at the Log-Likelihood Ratio Test.
   * Log-Likelihood Ratio Test
     + Log-likelihood NBD regression versus Poisson Regression = 4723.946
     + Log-likelihood NBD model versus Poisson model = 4797.154
     + Log-likelihood NBD regression versus NBD model = 18.104
   * Conclusion: Based on the AIC, BIC, and Log-Likelihood calculations, we would recommend using the NBD Regression as it shows an increase in performance against the NBD model and has low AIC and BIC among all of the models that were created

**Summary:** In this project, we developed and evaluated four different models – Poisson, NBD, Poisson Regression, and NBD Regression – using MLE to assess their effectiveness in predicting the number of purchases. We used key metrics like AIC, BIC, and the Log-Likelihood Ratio test to compare the different models and decided that the NBD Regression was the best model in predicting the number of books a given customer may buy. While some of the models performed well without the predictor variables, it is clear through these metrics that utilizing predictor variables in our model building improved the performance and will be more beneficial in prediction for Barnes and Noble. In our final NBD Regression model, the most impactful predictor variables appeared to be household size, age, and income. When each were increased the rate of purchases tended to increase. We recommend that Barnes and Noble focus on the demographics of older people with a larger household and are within a higher income bracket.