1. Model **initial_model** with all predictor variables (only main effects, no interactions), Poisson with log link
2. Initial checks of **initial_model**
    a. Goodness-of-Fit
        i. Note that since data is grouped, I can perform a Goodness-of-Fit test
        ii. I first checked expected counts, and there are none that are less than 1, and only about 5% (which is less than 20%) that are less than 5, thus the deviance goodness-of-fit and likelihood-ratio test statistics will be well described by a chi-square distribution.
        iii. Goodness-of-Fit test on **initial_model**. Failed with p-value = 4.278675e-34
            1. There is extremely strong evidence that the poisson model does not fit the data well.
    b. Overdispersion
        i. check_overdisperion() on **initial_model** -> Overdispersion detected
            1. There is overdispersion detected in the poisson model.
    c. Multicollinearity
        i. VIFs of **initial_model**
            1. All VIFs are less than 5, so there is no evidence of multicollinearity in the data.
    d. Influence
        i. Check Outliers (threshold = abs(studentized residuals) > 3))
            1. Obs: 1, 9, 12, 24, 27, 32, 41, 62, 75, 76, 77, 82, 89, 91, 97, 107
        ii. Check Leverage (threshold = hat value > 3 * (5) / 110)
            1. Obs: 41, 43, 50, 89, 94
        iii. Check Influence (threshold = Cook's Distance > F(0.5, 5, 105)
            1. None
        iv. Fit model **initial_model_wo_influential** without the outliers + high leverage points all predictor variables (only main effects, no interactions), Poisson with log link
            1. Comparing the **initial_model** fitted on the entire dataset to the **initial_model_wo_influential** fitted on the dataset with the 19 outliers or high leverage points removed, the coefficients themselves do not change much and the p-values/significance of the predictors do not change. However, the AIC does dramatically decrease once the 19 outliers and high leverage points are removed. However, since both of these models are fitted with Poisson which shows overdispersion, once I address the overdispersion issue, this might reduce extreme outliers' influence. Thus, I will not be removing these 19 observations just because

they are outliers or have high leverage and they are not influential. Since each row represents a unique neighborhood, dropping them would mean losing valuable data about store visits from those areas. However, it is good to acknowledge that these outliers and high leverage observations exist.

e. Since overdispersion was detected in the poisson model, I will now try to fit a negative binomial model.

3. Model **nb_inital_model** with all predictor variables (only main effects, no interactions), Negative Binomial with log link
4. Initial checks of **nb_inital_model**
    a. Goodness-of-Fit
        i. Note that since data is grouped, I can perform a Goodness-of-Fit test
        ii. I first checked expected counts, there are none that are less than 1, and only about 5% (which is less than 20%) that are less than 5, thus the deviance goodness-of-fit and likelihood-ratio test statistics will be well described by a chi-square distribution.
        iii. Goodness-of-Fit test on **nb_inital_model**. Passed with p-value = 0.2737936
            1. There is insufficient evidence that the negative binomial model does not fit the data well.
    b. Overdispersion
        i. odTest() on **nb_inital_model** -> p-value < 2.2e-16
            1. Reject H0 -> extremely strong evidence of overdispersion in the Poisson model, NB2 fits the data better than Poisson
    c. Multicollinearity
        i. VIFs of **nb_inital_model**
            1. All VIFs are less than 5, so there is not evidence of multicollinearity in the data.
    d. Influence
        i. Check Outliers (threshold = abs(studentized residuals) > 3))
            1. Obs: None
        ii. Check Leverage (threshold = hat value > 3 * (5) / 110)
            1. Obs: 15, 94
        iii. Check Influence (threshold = Cook's Distance > F(0.5, 5, 105)
            1. None
        iv. The negative binomial model which helps with overdispersion also reduces extreme outliers' influence. Before there were 19 outliers or high leverage points in the initial Poisson model. With the inital negative binomial model, there are only 2 high leverage points. There are no

outliers, and no points that are influential. Thus, since each row represents a unique neighborhood, dropping them would mean losing valuable data about store visits from those areas. However, it is good to acknowledge that these 2 high leverage observations exist.

e. Note: Did not fit ZIP or ZINB models because there is only one observation that has zero customer (response) so zero-inflated model would not be appropriate.
f. Negative binomial model checks out, so moving forward with that.

5. Stepwise regression (both directions) on all predictors and 2-way interactions using BIC
    a. I will first center all my predictors before running my stepwise regression since they are all quantitative and I will be trying to fit interactions, so I will center them to reduce multicollinearity.
    b. Started with model **nb_int**, which includes all predictors and 2-way interactions, Negative Binomial with log link
    c. Model **nb_step**, the result of the stepwise regression, contains the following terms (Negative Binomial with log link): units_c, income_c, compdist_c, storedist_c
        i. After adjusting the p-values using FDR, all the predictors in the model from stepwise are still significant using a 1% significance level.
    d. LRT between **nb_step** and **nb_int**
        i. p-value = 0.8882921
        ii. There is insufficient evidence that adding the extra parameters (in full model with all 2-way interactions) significantly improves the model fit compared to the candidate model. Thus, the simper (candidate) model is sufficient.
    e. AIC and BIC of **nb_step** and **nb_int**
        i. AIC
            1. **nb_step**: 691.8929
            2. **nb_int**: 708.1234
        ii. BIC
            1. **nb_step**: 708.0957
            2. **nb_int**: 754.0315
        iii. The candidate model (**nb_step**) has lower AIC and BIC values compared to the full model with interactions (**nb_int**).

6. Final checks of candidate final model **candidate_model**
    a. Model **candidate_model** contains the following terms (Negative Binomial with log link): units_c, income_c, compdist_c, storedist_c

i. Note: **candidate_model** = **nb_step**
ii. Although there are no interactions in the candidate model, I will still use the centered predictor values instead of the original predictor values because it doesn't change the predictor coefficients or interpretations, but it makes the intercept interpretation more applicable.

b. Goodness-of-Fit
   i. Note that since data is grouped, I can perform a Goodness-of-Fit test
   ii. I first checked the expected counts, there are none that are less than 1, and only about 5% (which is less than 20%) that are less than 5, thus the deviance goodness-of-fit and likelihood-ratio test statistics will be well described by a chi-square distribution.
   iii. Goodness-of-Fit test on **candidate_model**. Passed with p-value = 0.2926006
      1. There is insufficient evidence that the negative binomial candidate model does not fit the data well.

c. Overdispersion
   i. odTest() on **candidate_model** -> p-value < 2.2e-16
      1. Reject H0 -> extremely strong evidence of overdispersion in the Poisson model, NB2 candidate model fits the data better than Poisson

d. Multicollinearity
   i. VIFs of **candidate_model**
      1. All VIFs are less than 5, so there is not evidence of multicollinearity in the data.

e. Influence
   i. Check Outliers (threshold = abs(studentized residuals) > 3))
      1. Obs: None
   ii. Check Leverage (threshold = hat value > 3 * (4) / 110)
      1. Obs: 15, 30, 94
   iii. Check Influence (threshold = Cook's Distance > F(0.5, 4, 106)
      1. None
   iv. Our negative binomial candidate model has the same two high leverage points as our initial negative binomial model in addition to observation 30. However, since these observations do not have high influence, I will not be removing them. Additionally, since each row represents a unique neighborhood, dropping them would mean losing valuable data about store visits from those areas. However, it is good to acknowledge that these 3 high leverage observations exist.
   v. Comparing the candidate model fitted on the entire dataset to the candidate model fitted on the dataset with the 3 high leverage points

removed, the coefficients themselves do not change much and the p-values/significance of the predictors do not change. The AIC does decrease slightly by about 15 when the 3 high leverage points are removed. Since the interpretations do not change much and the overall results do not change, there isn't much to do except acknowledge that these high leverage observations exist. Since each row represents a unique neighborhood, dropping them would mean losing valuable data about store visits from those areas.

f. No problems in **candidate_model** to fix

7. Interpretation of the effects of all significant predictors in the final model (**candidate_model**)
8. Create a table of predictions (counts) from final model (**candidate_model**) for all combinations of predictors (min, mean, max) since all predictors are quantitative ->3^4 = 81 combinations of the predictors
9. Comment on unresolved problems
    a. 3 high leverage observations (Obs: 15, 30, 94)