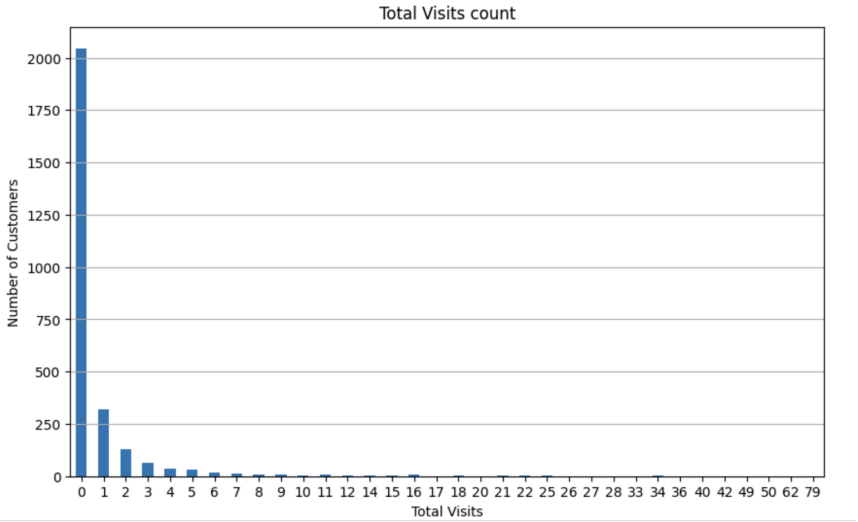
[i] First aggregate the data at total visit level and count the number of costumers. Then plot a bar chart. Show your plot and comment on what the distribution looks like. [5]



This looks like a poisson distribution, with a lot of 0, so zero inflated poisson model would be a better choice.

[ii] Write down a Poisson model that estimates the impact of covariates ((income, sex, age, and size) on number of visits. Show the equation, estimate GLM model, show the output and interpret the estimates. [10]

$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{income} + \beta_2 \times \text{sex} + \beta_3 \times \text{age} + \beta_4 \times \text{size}$$

Generalized Linear Model Regression Results

Dep. Variable:	Total_Visits	**No. Observations:**	2728
Model:	GLM	**Df Residuals:**	2723
Model Family:	Poisson	**Df Model:**	4
Link Function:	Log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-6291.5
Date:	Fri, 23 Feb 2024	**Deviance:**	10745.
Time:	05:26:04	**Pearson chi2:**	4.10e+04
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.06187
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.1262	0.406	-7.702	0.000	-3.922	-2.331
Income	0.0938	0.034	2.730	0.006	0.026	0.161
Sex	0.0043	0.041	0.104	0.917	-0.076	0.084
Age	0.5882	0.055	10.746	0.000	0.481	0.696
Size	-0.0359	0.015	-2.349	0.019	-0.066	-0.006

-

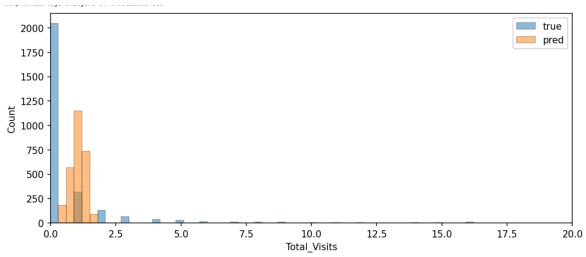
- $$\log(\lambda_i) = -3.1262 + 0.0938 \times \text{income} + 0.0043 \times \text{sex} + 0.5882 \times \text{age} - 0.0359 \times \text{size}$$
- **intercept (-3.1262)**:The intercept represents the log of the expected total visits count when all the predictor variables are 0.
- **income (0.0938)**: One unite increase in imcome, changes the log of the expected total visits count by 0.0938. The coefficient is significant(pvalue 0).
- **sex (0.0043)**: Being in male, compared to the female, changes the log of the expected total visits count by 0.0043. The coefficient is not significant(pvalue 0.917).
- **age (0.5582)**: One year older, changes the log of the expected total visits count by 0.5582. The coefficient is significant(pvalue 0.0).
- **Size (-0.036)**: One unit increase in size, changes the log of the expected total expected total visits count by -0.036. The coefficient is significant(pvalue 0.019).

[iii] You believe that income will go up by 10 percent in the coming year. How will it change the number of visits? We aggregated the data at visit level in part (i). Calculate the updated number of visits and plot the bar chart. Comment on what you see. [15]

Due to the positive coefficient of income, the number of visits will go up.

```
df['Income_1.1'] = df['Income'] * 1.1

df['Predicted_Visits'] = poisson_model.predict(df.assign(Income=df['Income_1.1']))
```



	ID	Total_Visits	Income	Sex	Age	Size	Income_1.1	Predicted_Visits
0	1	0	11.379394	1	3.871201	2	12.517333	1.294227
1	2	5	9.769956	1	4.043051	1	10.746952	1.257091
2	3	0	11.082143	0	3.332205	2	12.190357	0.910201
3	4	0	10.915088	1	3.951244	3	12.006597	1.247535
4	5	0	10.915088	1	2.833213	3	12.006597	0.646288
...	...	...	...	...	...	...	...	...
2723	2724	0	9.528794	1	2.944439	2	10.481674	0.619864
2724	2725	0	11.379394	0	3.970292	2	12.517333	1.366079
2725	2726	0	11.191342	1	3.044522	3	12.310476	0.752996
2726	2727	0	10.532096	1	2.890372	4	11.585306	0.619821
2727	2728	0	11.736069	1	2.833213	3	12.909676	0.703437

2728 rows × 8 columns

A lot of customers are now predicted to have higher counts of total visits instead of 0, after we increase income by 10%. So income increase indeed increases total count of visits.

Explain the mechanism behind zero inflated model. What modification to the Poisson model should be made? Explain clearly. [5]

Zero-inflated model is used when you believe there are a lot of data points who got 0 count but it was just because they did not even performed the action you are interested in counting the results. And in order to differentiate between people who got 0 count and those who did not even take action. We use zero-inflated model to include a probability estimation of they being in which group.

1. first (model 1), use log of income as a covariate for zero inflation part and the other two as covariates for Poisson model,

```
zip_training_results = sm.ZeroInflatedPoisson(endog=y, exog=X, exog_infl=data[['log_income']], inflation='logit').fit(maxiter=100)
print(zip_training_results.summary())
```

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
inflate_log_income     0.0036      0.001      4.986      0.000       0.002       0.005
Intercept              0.8156      0.062     13.123      0.000       0.694       0.937
gender                 0.0015      0.008      0.181      0.856      -0.015       0.018
log_income             0.0078      0.005      1.535      0.125      -0.002       0.018
==============================================================================
```

2. and second (model 2), use Gender as covariate for zero inflation part and the other two for Poisson model. [5]

```
expr = 'number_calls ~ gender + log_income'
y,X = dmatrices(expr, data, return_type='dataframe')
zip_training_results = sm.ZeroInflatedPoisson(endog=y, exog=X, exog_infl=data[['gender']], inflation='logit').fit(maxiter=100)
print(zip_training_results.summary())
```

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
inflate_gender     0.0500      0.013      3.962      0.000       0.025       0.075
Intercept          0.8154      0.062     13.098      0.000       0.693       0.937
gender             0.0064      0.008      0.767      0.443      -0.010       0.023
log_income         0.0074      0.005      1.466      0.143      -0.003       0.017
==============================================================================
```

Can you explain what the economic meaning of these covariates is? Plot the actual and predicted values. Which model (model 1 or 2) seems like a better fit? [5]

Inflate_log_income. One percent change in income will increase the probability of not being the ‘true zeros’.

Inflate_gender. Being female, compared to being male will increase the probability of not being the ‘true zeros’.

Gender. Being female, compared to being male increases the expected log count of number of calls.

Log income. One percent change in income will increase the expected log count of number of calls.

The models look pretty much the same, the second one seems to have smaller p values, also from the predictions, we can see the predictions are all around 1.2, which says our models are not learning the relationship between gender, log income and number of calls