1. **Introduce a business problem, questions, issues, or elements for investigation**

Our main business problem we are looking to address is how household factors such as marriage, childcare responsibility, and household structure affect and influence self-employment status. To execute this, we plan to investigate the CPS labor force data set and select variables that represent household situations.

2. **Discuss the data your team selected and reasoning for why it can be used to address the original business problem (including format, number of records, level or units of the data, variables, and any filtering or adjustments that was done)**

We started by gathering data covering a ten-year span. Using SQL, we combined these tables and created a dataset with over 3,443,020 records. Using the CPS Labor Force dataset, we decided to select a few variables including hhid, hhid2, lineno, married, age, ch02, ch05, ch35, ch613, ch1417, hhnum, hoh79, selfemp, selfinc, and ownchild. These fields were chosen to capture both household-level and individual-level characteristics that may influence self-employment patterns, allowing us to examine how family structure and household composition intersect with self-employment, and provide a comprehensive view of the socio-economic factors at play.

For our data cleaning, We started off by combining hhid and hhid2 and created a new primary key called hhhid_full. We merged selfemp and selfinc variables as we are only interested in whether or not the person is self employed. To do this, we had to drop NAs because NA cannot be an integer and we are looking for 0s and 1s. Afterwards, we reset the index. For children age, variables ch02, ch35, and ch05 were redundant variables so we removed ch02 and ch35. Our assumption was that a child that is 0-2 years old and 3-5 years old is also 0-5 years old. We then replaced the null values with the Mode for variables ownchild, ch05, ch613, and

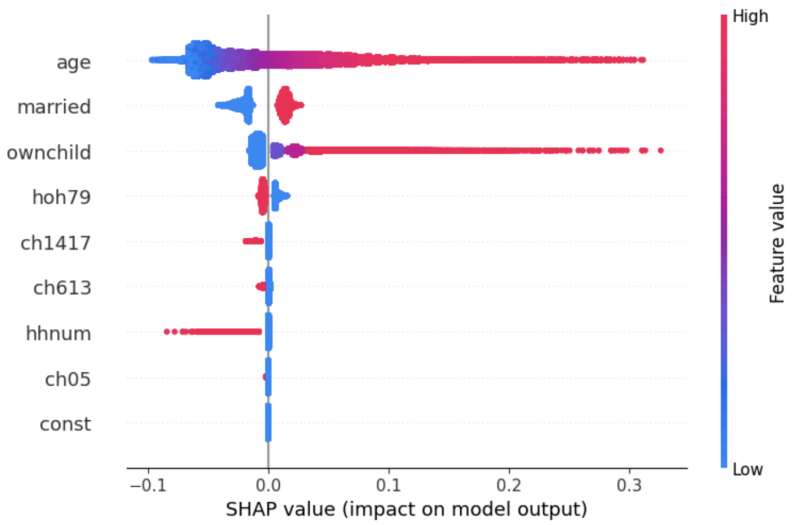
ch1417. The data was then filtered for age greater than or equal to 18 because our assumption is that you need to be at least 18 years old to work a proper corporate job without restrictions. We converted the fields to numerical data types except for the primary key. After this process, we were left with 2,212,505 records.

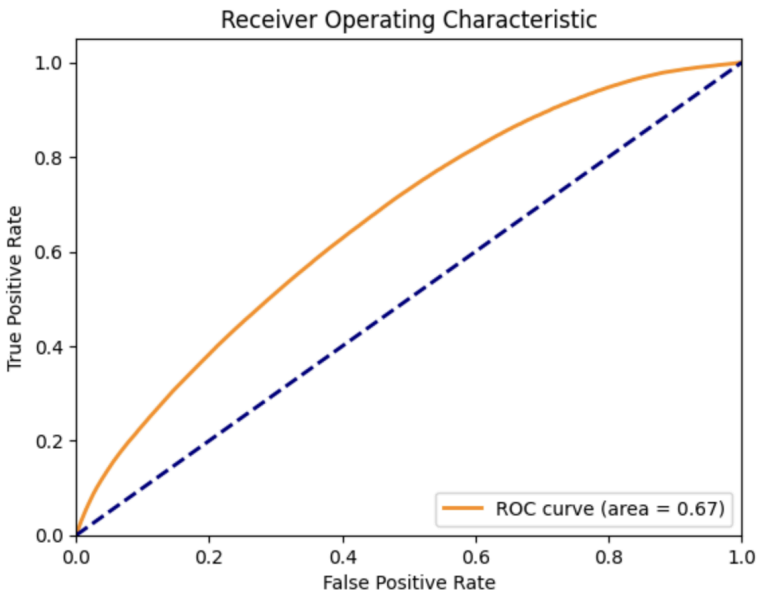
3. **Detail any modelling and analyses that were performed including an assessment of the accuracy of your model results (although Tableau has some Analytics functionalities this may be done outside of Tableau and a summary of the model or outputs can be imported for inclusion in your visualizations)**

Query job e2e80702-c0d8-48d1-b91e-feb660db3524 is DONE. 338.8 MB processed. Open Job

```
Optimization terminated successfully.
         Current function value: 0.320919
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:        self_employment   No. Observations:              2212505
Model:                         Logit   Df Residuals:                  2212496
Method:                          MLE   Df Model:                            8
Date:                Mon, 14 Apr 2025   Pseudo R-squ.:                0.05117
Time:                        00:36:30   Log-Likelihood:           -7.1003e+05
converged:                       True   LL-Null:                  -7.4833e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.1139      0.016   -263.298      0.000      -4.145      -4.083
age            0.0410      0.000    221.499      0.000       0.041       0.041
ch05          -0.0120      0.010     -1.155      0.248      -0.032       0.008
married        0.4026      0.008     50.128      0.000       0.387       0.418
hhnum         -0.1464      0.011    -13.395      0.000      -0.168      -0.125
ch613         -0.0517      0.010     -5.158      0.000      -0.071      -0.032
ch1417        -0.1114      0.009    -12.097      0.000      -0.130      -0.093
hoh79         -0.1393      0.009    -15.613      0.000      -0.157      -0.122
ownchild       0.1690      0.005     32.335      0.000       0.159       0.179
==============================================================================
```

```python
# independent and dependent variables
independent_vars = [
    'age', 'ch05', 'married','hhnum',
    'ch613', 'ch1417',
    'hoh79', 'ownchild'
]
dependent_var = 'self_employment'


df = df.to_pandas()

for col in [dependent_var] + independent_vars:
    mode = df[col].mode()
    if not mode.empty:
        df[col] = df[col].fillna(mode[0])


X = df[independent_vars].astype(float)
X = sm.add_constant(X)
y = df[dependent_var].astype(float)


logit_model = sm.Logit(y, X)
result = logit_model.fit()


print(result.summary())
```






```
Pseudo R-squared (McFadden's R-squared): 0.05117490441009054
Accuracy: 0.8939202397282718
Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94   1977816
           1       0.24      0.00      0.00    234689

    accuracy                           0.89   2212505
   macro avg       0.57      0.50      0.47   2212505
weighted avg       0.82      0.89      0.84   2212505

Confusion Matrix:
[[1977797      19]
 [ 234683       6]]
```

After running a logistics regression analysis, we see that age is significant and that older individuals are more likely to be self-employed. Ch05 is not significant with a p-value of 0.248. Having children under 5 does not significantly affect self-employment. Married status is significant and married individuals are more likely to be self-employed. Hhnum is significant and the larger the household size, the less likely it is for the individual to be self-employed. Ch613 is significant and having children aged 6-13 is negatively associated with

self-employment. Ch1417 is significant as well and having children aged 14-17 is negatively associated with self employment. Hoh79 is actually negatively associated with self-employment in our model. Ownchild is significant and having any owned children increases likelihood of self-employment.

Our major takeaways include having children lowers the rate of self employment. Older individuals are more likely to be self-employed. The Head of Household is less likely to be self-employed. Finally, owning a child increases the likelihood probably only if the child is over 18.

This gives out the results of our hypothesis testing. Our first hypothesis is married individuals are more likely to be self-employed than unmarried individuals. Based on our regression, we reject the null hypothesis as there is evidence that married individuals are more likely to be self-employed. Our second hypothesis is that individuals with children are more likely to pursue self-employment due to flexibility in the schedule. Based on our regression, we reject the null hypothesis as there is evidence that suggests that individuals with school-aged children are less likely to be self-employed but individuals that own children are more likely to be self-employed. Our last hypothesis is that household heads or primary earners are more likey to be self-employed than other household members. Based on our regression we reject the null hypothesis as there is significant evidence, but in the opposite direction. Household heads are less likely to be self-employed.

For further model analysis we did a SHAP analysis. The results showed that Age has the strongest influence on the model's prediction. Married and ownchild are also impactful variables. Overall, older age positively impacts the prediction outcome. Married status and child ownership also affect predictions but with a more mixed influence.

In terms of Model Accuracy, we first ran an ROC chart. Our AUC score came out to 0.67 which means the model has limited ability to distinguish between classes. A perfect score would be a 1.0 and a 0.5 would mean random guessing. A score of 0.67 shows that the model performs slightly better than random but there is room for more improvement. Next, we ran a confusion matrix with an accuracy of 89.39%. This accuracy however, is misleading due to class imbalance. For class 0, the precision and recall are both high. For class 1, there is a recall of 0.00 meaning the model fails to detect positives. Overall our model is biased towards the majority class (class 0). It predicts class 1, which is self-employed, very poorly and makes it unsuitable for applications where identifying self-employment is critical.

4. **Present the visuals and dashboard you've created and discuss why what you selected to show is useful in the context of the business**

For our dashboard and visuals we used python and ran a command in the prompt. We started off by looking at age and its relationship with self-employment. We used a line chart to show the percentage of self-employment rate for each age. There is a positive and upwards trend showing that the older the individual is, the more likely they are to be self-employed.

For children's age, we used bar graph distribution to show the effect on self-employment. Through our distribution it is shown that those who do not have children in the school age are more likely to be self-employed and those who do have children are less likely to be self-employed.

For our married variable we created a bar chart distribution. The bar chart shows the proportion of self-employed individuals within married people seems somewhat higher compared to unmarried people, indicating marriage might have a slight positive impact on

self-employment. For those not married (0), most individuals are not self-employed. There's a smaller proportion of self-employed people relative to married individuals. Overall, while most individuals are not self-employed regardless of marital status, married people appear slightly more likely to be self-employed than unmarried ones.

For hhnum, we used a line chart to show the percentage of self-employment depending on the number of people in a household. Our line chart explains that there is a higher self-employment percentage with 1 person in the household and then it trends downwards as the number of people increases. Once it hits 8, the number goes back up.

For hoh79, our regression initially shows a -0.012 coefficient while holding all other variables constant. However, our visualizations show that there is actually a positive influence of being head of household on self-employment. This is because our distribution and line graph is just a trend and is not used for predictive purposes. Our visualizations tell how outcomes differ by a variable while logistic regressions tell how a variable really impacts the outcome while holding others constant.

Finally, for ownchild, the more children you own the more likely you are to be self-employed. Considering that owning a school-age child is negatively correlated, we are under the assumption that this is for owning a child over 18. There is a positive trend that the more children you own the more likely you are to be self-employed until you own over 10 children. Over 10 is when the trend goes downwards.

Our correlation matrix shows how numerical variables are correlated with each other. The color scale goes from -1 (strong negative) to +1 (strong positive). For age vs. self-employment: There is a positive correlation, indicating that older individuals are more likely to be self-employed. For ownchild vs. self_employment: The correlation appears positive but weak.

For ch05, ch613, ch1417, there is a slightly negative correlation, suggesting that individuals with younger children are less likely to be self-employed. For married vs. self-employment: Slight positive correlation, suggesting married individuals may have a slightly higher tendency to be self-employed. In conclusion, age and number of children have some influence on self-employment, with certain older age groups showing notably high self-employment rates, especially when there are more children. The correlation matrix supports this pattern but also shows that these variables are only moderately correlated with self-employment. Overall, age appears to be the most influential numerical variable related to self-employment in this dataset.

Based on our model, the SBA should focus on targeting: Older individuals, families that own a child over 18 and avoid families with children that are still of school age, married individuals, non head of household members might be multivariate meaning it can have more than one possibility. Head of household might not be a good factor to target.

5. **Conclude with what would be needed to continue to update your analyses and identify any deficiencies or next business steps.**

Based on our current model, we suggest that we need to update the dataset regularly with the most recent demographic and employment trends to account for changes in the workforce, including shifts in marital status, family structure, self-employment, and other variables. Then incorporate external data sources such as government labor statistics, industry-specific reports, or surveys to provide a broader view of the labor market trends and changes. We should also focus on emerging factors such as the growing influence of the gig economy, remote work trends, and shifts in employee values and preferences regarding work-life balance. Finally, examine behavioral data such as job satisfaction, career progression, and aspirations to capture deeper

insights into the motivations behind self-employment and its connection to traditional employment roles. For further analysis we should focus on exploring other significant predictors to see their impact on self-employment as well as testing different economic conditions or regional variables that may influence self-employment.

As for deficiencies and next steps, our model is imbalanced and not effective for detecting the minority class. We can try resampling using Synthetic Minority Oversampling Technique to reduce the chance of oversampling, The pseudo R-squared is low (around 0.05), indicating that the model explains very little of the variation in the data. The features we selected may not be strong predictors in their current form or need further transformation or engineering. To do this we can create new variables or interactions based on domain knowledge, normalize or scale numerical variables, and consider non-linear transformations if needed. Overall, the current model is not reliable enough for identifying self-employed individuals - which may affect targeting strategies, policy design, or financial planning. Improving the model can help in better segmenting customer groups, tailoring financial services, or support programs and informing future data collection needs.