Dataset: https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction

Timeline

04/28:

- Zoom meeting:
  - Discuss expectations of the project, overall framework and timelines
  - Priority tasks:
    - Strong EDA, each person conducting their own
    - Addressing all required points in the rubric
  - Turner has been notified
    - If he says no or there are issues, go with 50k dataset

03/31:

- Meet with Turner and double check if he hasn't responded regarding our dataset
- Make final decision on the dataset
- Start EDA

04/03

- Meet to discuss findings of our EDA (each person performs)
- Discussion:
  - Adam checked glm assumptions (multicolinearity and VIFs)
  - Adam condensed datapoints by sector, education, and CONTINENT of origin
  - Relationships could be designated as relation within the household
- Tasks:
- Start on Objective 1
  -
- Build a general framework for our final EDA:
  - Start with treemap by country
  - Rename all columns:

```
colnames(adult) <- c('age', 'workclass', 'similar_pop_count', 'education_level',
                     'years_education', 'marital_status', 'occupation', 'household_role', 'race',
'gender',
                     'capital_gain', 'capital_loss', 'hours_per_week', 'native_country', 'income')
```

  - Convert from country to continent, and the explanation for how and why each country was classified accordingly

```
# Define the new levels for native_country
```

```
new_levels <- list(
  EAsia = c("Vietnam", "Laos", "Cambodia", "Thailand", "China", "Hong", "Taiwan",
"Philippines", "Japan"),
  SAsia = c("India", "Iran"),
  NorthAmerica = c("Canada", "Mexico", "United-States"),
  CentrAmerica = c("Cuba", "Dominican-Republic", "Guatemala", "Haiti", "Honduras",
"Jamaica", "Trinadad&Tobago", "Nicaragua", "El-Salvador", "", "?"),
  SouthAmerica = c("Ecuador", "Peru", "Columbia", "South"),
  Europe = c("France", "Germany", "Greece", "Holand-Netherlands", "Italy", "Hungary",
"Ireland", "Poland", "Portugal", "Scotland", "England", "Yugoslavia"),
   USTerritory = c("Outlying-US(Guam-USVI-etc)", "Puerto-Rico")
)
```

- Condense workclass category:

```
new_levels <- list(
  other = c("", "?"),
  clerical = c("Adm-clerical"),
  midskill = c("Craft-repair", "Machine-op-inspct", "Transport-moving"),
  lowskill = c("Handlers-cleaners", "Other-service", "Priv-house-serv", "Armed-Forces"),
  highskill = c("Sales", "Tech-support", "Protective-serv", "Prof-specialty",
"Exec-managerial"),
  agriculture = c("Farming-fishing")
)
```

- Condense marital status category:

```
# The new marital status categories
new_levels <- list(
  divorce = c("Divorced", "Separated"),
  married = c("Married-AF-spouse", "Married-civ-spouse", "Married-spouse-absent"),
  notmarried = c("", "Never-married"),
  widowed = c("Widowed")
```

- Condense education level:

```
# The new education level categories
new_levels <- list(
  no_edu = c("", "Preschool"),
  primary = c("1st-4th", "5th-6th"),
  secondary = c("7th-8th"),
  highsch = c("9th", "10th", "11th", "12th", "HS-grad"),
  assoc = c("Assoc-acdm", "Assoc-voc", "Some-college"),
  undergrad = c("Bachelors", "Prof-school"),
  master = c("Masters"),
  phd = c("Doctorate")
)
```

- Convert all missing variables in the same way (with mode of that variable)
- Split our Test and Train sets the same:
    - 80:20

```
Set a seed for reproducibility
set.seed(1234)

# Calculate the number of rows to select
sample_size <- round(0.8 * nrow(adult))

# Create a random sample of indices
index <- sample(1:nrow(adult), size = sample_size, replace = FALSE)

# Create the training set using the selected indices
training_set <- adult[index, ]

# Create the testing set with the remaining indices
testing_set <- adult[-index, ]
nrow(training_set_LR)
nrow(testing_set_LR)

```

    - Make a new column from capital gains and capital loss: invested. Yes or no, depending if they have a 0 or EITHER a gain or a loss

04/06

- Meet to discuss findings for Objective 1
- Start on Objective 2
    - Adam: random forest
    - Steven: PCA?
    - Joel: LDA? and QDA?

Notes:

- For objective 1, make sure you have PCA done for your model to see if you can simplify the model in any way
- Fine-tune your final models from Objective, check that assumptions for logistic regression are met
- Look for outliers that might need to be removed
- Show attempts to add complexity:
    - Show that adding multiplicative interaction term didn't help

        - Make sure we have slides that iterate that added complexity did not help the model
        - Explain that altering threshold optimized the model we did choose

- Talk with Turner: Can we condense our data set by changing factor levels for interpretability (education level and continent, for example
- Show performance doesn't change much when we use untransformed data (justification for using our transformed dataset)
- Each person should take their final model and try adding complexity. Show that the complexity helps or doesn't make a difference
- Once you have your final model, build your interpretation of the coefficients with confidence intervals (written interpretation is required; just showing the table of coefficients is not enough. State the findings so there is no confusion)
- Add EDA and Objective 1 findings to Powerpoint Draft so we can talk with Turner
- Start on objective 2:
    - Build a predictive model for the purposes of prediction, not interpretability (so PCA should be possible)
        - Report performance metrics: Sensitivity, Specificity, Prevalence, PPV, NPV, and AUROC, threshold
- Have records of all your code so we can build an RMD file

04/10

- Meet to discuss findings for Objective 2
- Start on conclusions, final report, and presentation

04/14

- Showcase to Dr. Turner
- Start refining final products

04/15

- Finish all deliverables, record video

04/17

- Have final video edited