This README file will help you understand how to run/use the technical implementation of our Data Science Project for INST737 - Income Analysis.

Python Code:
To run this code, you will need to open the file in Google Colaboratory.

No software installation is required to be done manually, there are a few libraries which will be imported once you run the code, as follows:

```
import pandas as pd
import numpy as np
import io
import seaborn as sns
import matplotlib.pyplot as plt
import re
import scipy as sp
import statsmodels.api as sm
import umap
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.cluster import KMeans
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree
Classifier
from sklearn.model_selection import train_test_split # Import
train_test_split function
from sklearn import metrics #Import scikit-learn metrics module for accuracy
calculation
from sklearn.tree import export_graphviz
from six import StringIO
from IPython.display import Image
import pydotplus
from sklearn.ensemble import RandomForestClassifier
from google.colab import files
```

Additionally, for clustering, we have used UMAP, which will also be installed to your Google Colaboratory Runtime once you run our code. (Code below)

```
!pip install umap-learn
```

To run this code,

Go to Runtime -> and click on Run all

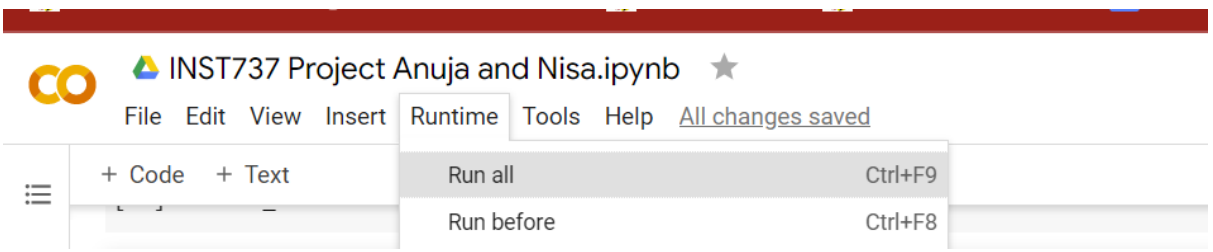
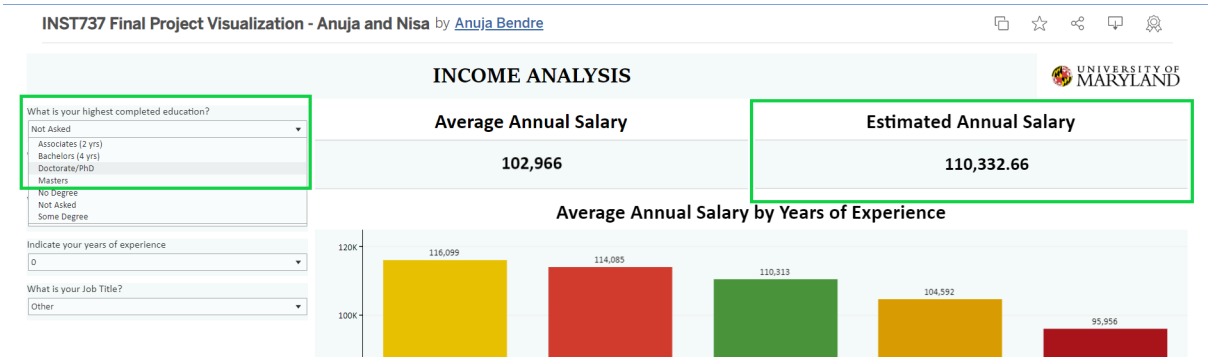


Tableau Public Dashboard:

To access the dashboard, click on the link below:
https://public.tableau.com/app/profile/anuja.bendre/viz/INST737FinalProjectVisualization-AnujaandNisa/Dashboard1?publish=yes

To predict the estimated annual salary based on factors such as Education, Employment Status, Gender, Years of Experience and Job Title, use the dropdown options on the left and wait 5-10 seconds for the value on the right to refresh based on the values provided.



Available factors:

Additionally, you can use the radio buttons on the left to check the observed annual average salary distribution based on each of the factors: (based on Source Data)