**Targets and Business Goals:**

As an XYZ Corporation data analyst, my primary aim is to develop a predictive model using the U.S. Census Bureau dataset, focusing on factors influencing individual income, particularly whether it exceeds $50,000. The overarching business objective for XYZ Corporation is to increase enrolments in their degree programs by refining marketing strategies based on the model's insights. This involves creating an accurate predictive tool, identifying key factors affecting income, and tailoring marketing efforts to attract individuals with a higher likelihood of exceeding the $50,000 income threshold. The ultimate goals are to boost enrolments, optimize marketing return on investment, and maintain a competitive edge in the education sector through effective data-driven strategies.

**Problem Statements:**

1. Developing a precise model to predict individual income levels based on demographic features, addressing the complexities of diverse datasets and potential biases.
2. Identifying the most influential factors from the Census data to tailor marketing strategies effectively, ensuring enrolments increases while avoiding overreliance on specific demographics.
3. Addressing issues of missing or inconsistent data in the Census dataset, requiring robust data cleaning and imputation strategies for accurate and reliable model development.
4. Bias Mitigation in Predictions: Tackling potential biases within the predictive model to ensure fair and unbiased outcomes, especially in sensitive areas such as income thresholds, to uphold ethical standards and inclusivity.

**Data Integrity Assumptions:**

***Representativeness***: Assuming that the dataset from the United States Census Bureau is a representative sample of the broader population, reflecting the diversity of socio-economic characteristics and demographics in the United States.

***Quality of Census Data****:* Assuming that the underlying data collected by the U.S. Census Bureau is accurate and reliable, with the appropriate measures taken to maintain data quality and integrity during collection and processing.

***Data Completeness*:** Assuming that the dataset is reasonably complete, with minimal missing values, and that any imputations or handling of missing data do not significantly impact the analysis.

***Temporal Relevance****:* Assuming that the dataset is relevant to the time under consideration, recognizing that socio-economic factors can change over time and may impact the generalization of the predictive model.

***Data Independence****:* Assuming that individual records in the dataset are independent, meaning that the characteristics of one person do not directly influence the characteristics of another person in the dataset.

**User Stories:**

***Business Analyst Stories:***

1. ***Income Predictor***

- Explore the correlation between education level and marital status.

- Determine the impact of the combination on an individual's income.

2. ***Gender-Based Income Disparity***

- Investigate the combination of gender and occupation.

- Identify potential gender-based income disparities.

3. ***Regional Income Variation***

- Explore the interaction between native country and work class.

- Discern regional income variations for targeted strategies.

4*.****Diversity Impact:***

- To analyze the correlation among race, occupation, and weekly working hours to gain insights into the impact on income diversity.

- Provide valuable information for promoting workplace inclusivity.

**Marketing User Stories:**

5. ***Targeted Marketing for Professions***

- Leverage the combination of education, occupation, and work class.

- Tailor marketing campaigns for specific professions.

6. ***Personalized Outreach for Marital Status***

- Analyse the relationship between marital status, race, and hours worked per week.

- Create personalized outreach strategies for diverse audience segments.

7. ***Cultural Tailoring***

- Understand the impact of native country, race, and education on income.

- Tailor promotional content for diverse cultural backgrounds.

8. ***Gender-Specific Program Promotion***

- Investigate the combination of gender, education, and capital gain.

- Create gender-specific marketing initiatives.

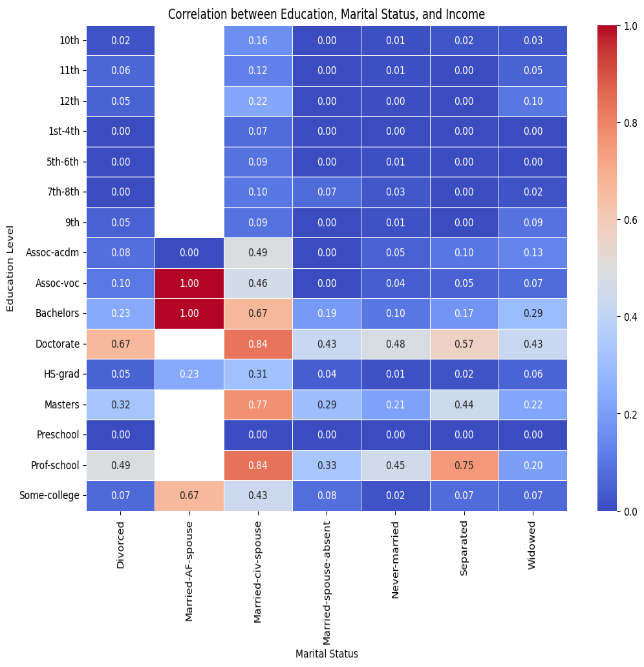
9. ***Regional Enrolments Campaigns***

- Explore the relationship between native country, work class, and education.

- Develop targeted enrolments campaigns for specific regions.

**Visualizations:**

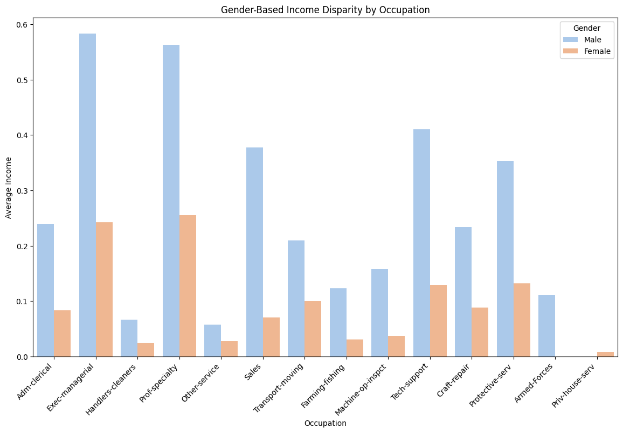
***1.* Income Predictor*(multivariate):***

The provided heatmap visualization illustrates the correlation between education level, marital status, and income, focusing on individuals who are married with an AF spouse and those in a civil partnership. The design process involved several steps: first, I converted the income categories into numeric values for analysis. Next, I created a pivot table to aggregate the average income for different combinations of education and marital status. The heatmap was chosen as it provides a clear and concise overview, allowing for easy identification of patterns and trends.

**Conclusion:**

The insight gained from the visualization suggests that individuals with Doctorate, Associate Vocational (Assoc-voc), Professional school, Bachelor's, and master's degrees, specifically those in a civil partnership, contribute more to the higher income bracket. This finding highlights the potential impact of education and marital status on income, providing valuable insights for targeted marketing strategies or program promotions tailored to these demographic segments.

2. **Gender-Based Income Disparity*(multivariate):***

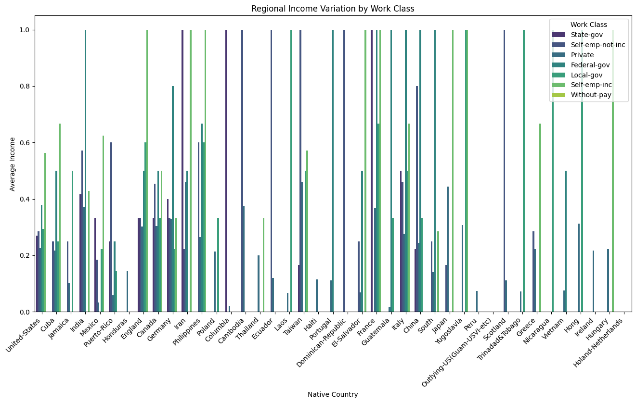
The grouped bar plot visualization demonstrates the correlation between gender, occupation, and income, specifically focusing on potential gender-based income disparities. The design process involved several steps: first, I created a numeric representation of income ('income\_numeric') where 1 represents income > $50,000 and 0 represents income <= $50,000. Then, a grouped bar plot was chosen as it effectively compares the average income for different occupations between genders. The x-axis represents various occupations, the y-axis represents the average income, and bars are grouped by gender. 

**Conclusion:**

The result of the visualization indicates that, on average, men contribute more to the higher income bracket (income > $50,000) compared to women across various occupations. This insight is crucial for understanding potential gender-based income disparities and can inform strategies aimed at promoting workplace equality and fairness.

**3.** **Regional Income Variation:**

The grouped bar plot shows the average income for each combination of native country, work class, and income. The 'income\_numeric' column, representing income > $50,000 (1) or <= $50,000 (0), is used .The visualization of regional income variation by work class and native country reveals insights into income disparities across different regions and work classes. The design process involved creating a numeric representation of income, using a grouped bar plot to compare average incomes. A pivot table was created to aggregate counts of occurrences for each combination of native country and work class.

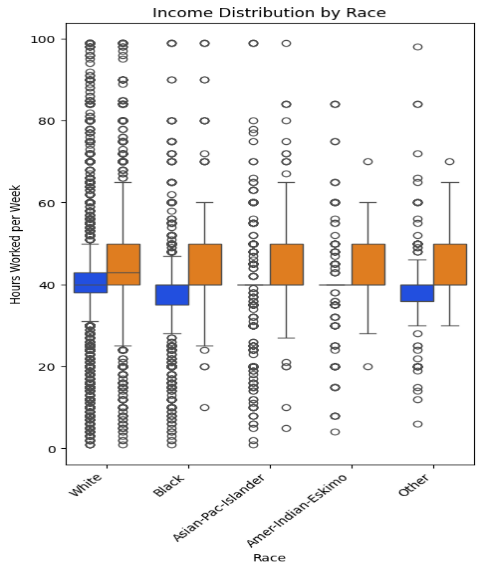


**Conclusion:**

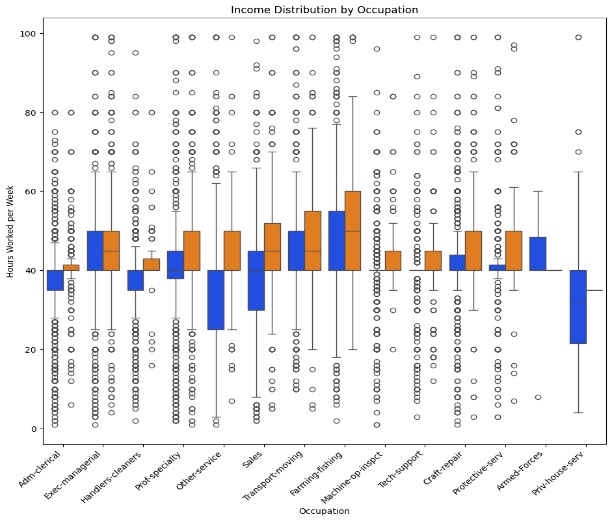
The visualization of top contributing countries for each work class, after filtering out entries from the United States, reveals international diversity within the dataset. Key insights include the Philippines being prominent in 'Federal-gov,' Mexico in 'Local-gov' and 'Private,' Cuba in 'Self-emp-inc,' India in 'State-gov,' and the Philippines again in 'Without-pay.' These findings inform targeted strategies for diverse recruitment, program promotions, and international initiatives within specific work classes.

**4.Diversity Impact***(****Univariates*):**

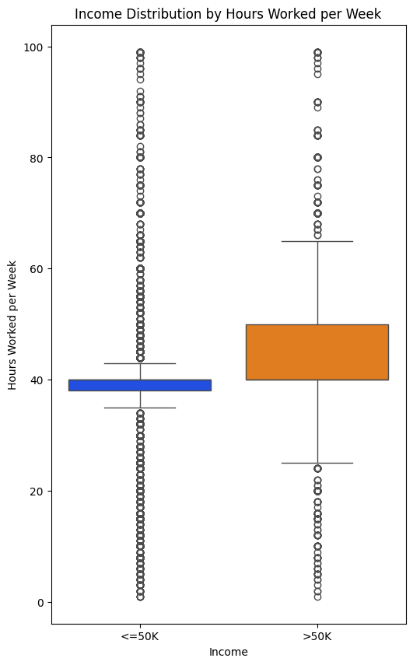
Univariate analysis involves examining a single variable at a time, and in the context of this user story, focusing on one variable (e.g., race, occupation, or hours worked per week) independently. This allows for a deeper understanding of the distribution.



The box plot helps understand the variation in hours worked per week for each race, showing the median, quartiles, and potential outliers. This information is crucial for UVW College to recognize patterns related to hours worked within different racial groups, assisting in tailoring marketing strategies to accommodate diverse schedules.



By examining the variation in hours worked per week within various occupations, UVW College gains insights into the work-hour patterns associated with specific jobs. This aids in shaping marketing efforts to address the lifestyle and time constraints of individuals in different occupations, facilitating targeted program promotions.



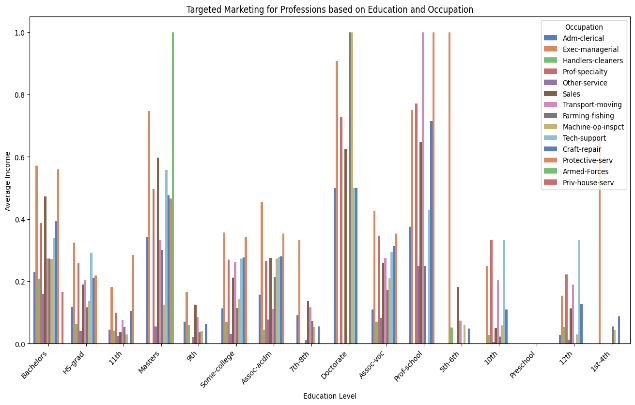
The box plot showcases how income levels are distributed across different ranges of hours worked per week. This insight is valuable for UVW College to identify income patterns concerning the time commitment of individuals. Understanding the income distribution helps in designing marketing strategies that align with the financial expectations of the target audience.

**Conclusion:**

Utilizing visualizations based on race, occupation, and weekly working hours offers crucial insights for UVW College's enrollment strategy. By understanding diverse factors influencing income levels, the college can tailor marketing efforts to meet specific needs, optimizing outreach and aligning strategies with the unique characteristics of the target population. Leveraging United States Census Bureau data contributes to a more precise and impactful enrollment strategy, poised to boost overall enrollment at UVW College.

**5.Targeted Marketing for Professions:**

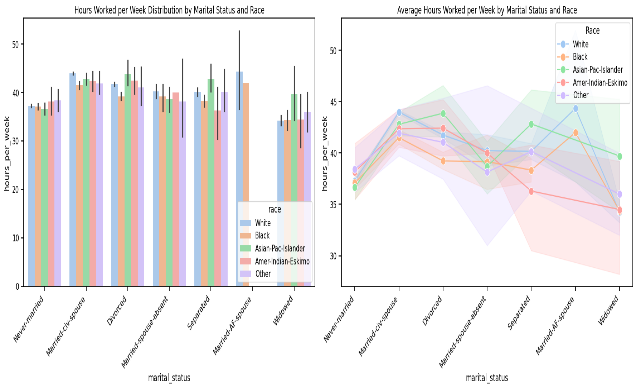
Tailoring the marketing campaigns for specific professions based on the combination of education, occupation, and work class, The grouped bar plot shows the average income for each combination of education, occupation, and work class. The 'income\_numeric' column, representing income > $50,000 (1) or <= $50,000 (0), is used for this visualization. The plot allows for the comparison of average income across different professions, guiding targeted marketing strategies based on education and occupation***.***



**Conclusion:**

Key observations include higher average incomes for individuals with a 'Doctorate' across various occupations, especially in 'Prof-specialty' and 'Exec-managerial.' Additionally, those with a 'Masters' or 'Bachelors' degree show significant income differences across professions, suggesting diverse income patterns. This insight guides targeted marketing efforts, emphasizing professions where higher education levels align with lucrative occupations, and enables strategic promotions and program offerings to attract individuals with specific educational backgrounds and career aspirations.

**6. Personalized Outreach for Marital Status*(multivariate):***

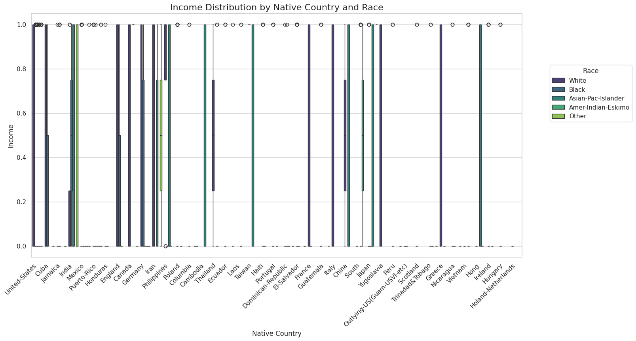
In the first visualization, a grouped bar plot was employed to illustrate the distribution of hours worked per week among various marital statuses, with each bar further divided by race. This allows for a comprehensive comparison of average working hours across different marital statuses and racial groups. The second visualization utilized a line plot to demonstrate the average hours worked per week for distinct racial groups within each marital status. The markers on the lines highlight specific data points, aiding in identifying trends. 

**Conclusion:**

Both visualizations aim to provide insights into the nuanced relationship between hours worked, marital status, and race. These analyses can be instrumental in informing UVW College's efforts to develop a predictive model for individual income determination based on key factors, ultimately supporting targeted marketing strategies to boost enrollment in their degree programs.

**7. Cultural Tailoring:**

The boxplot visualizes how income varies across different native countries. UVW College can tailor promotional content specific to each country, considering income disparities and cultural nuances. The boxplot differentiates income distribution by race, providing insights into how race influences earnings. UVW College can use this information to create culturally sensitive marketing materials that resonate with diverse racial backgrounds.The visualization can also reveal how education levels impact income. Understanding the educational background of individuals within each native country and race category is crucial for crafting targeted content.



**Conclusion:**

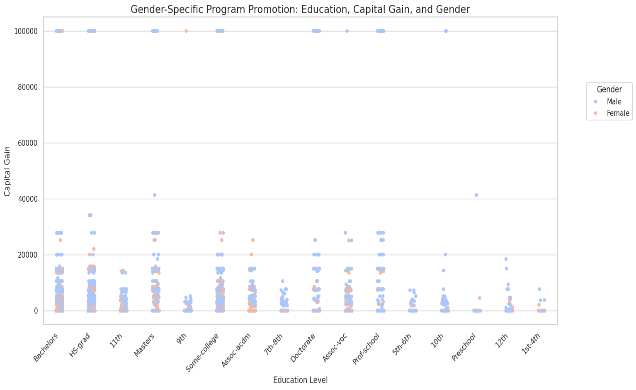
UVW College can leverage the identified patterns to tailor promotional content that aligns with cultural backgrounds. For example, campaigns can be designed to highlight specific degree programs based on income trends within particular demographics.Enrollment Boost Strategies: By tailoring marketing efforts according to the identified factors, UVW College can strategically promote degree programs that align with income trends. This targeted approach is likely to boost enrollment, aligning with the business objective.

**8.** **Gender-Specific Program Promotion*(Bivariavte):***

The strip plot visualizes the combination of gender, education level, and capital gain.

The plot categorizes data points based on the education level. The vertical position of each point represents the capital gain associated with the different colors distinguish between genders, with each point belonging to either the 'Male' or 'Female' category. Corresponding education level. The plot allows us to observe how capital gain varies across different education levels. For each education category, the distribution of capital gain can be assessed.

By utilizing different colors for each gender, the plot provides insights into how males and females contribute to capital gain within each education level. It helps identify any gender-specific trends or disparities.

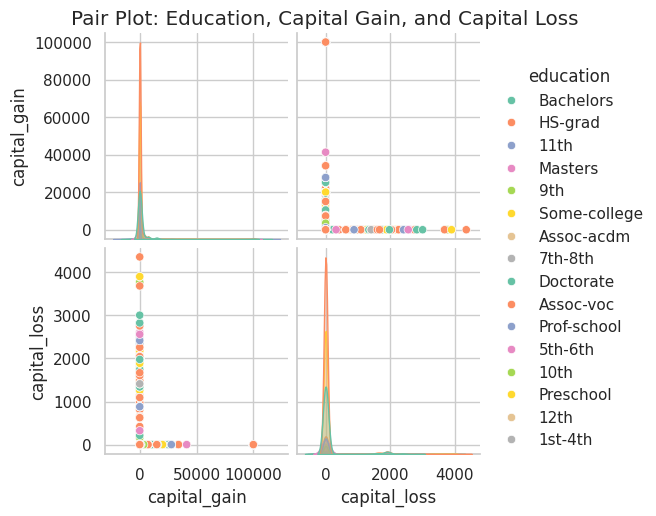


**Conclusion:**

The plot can reveal whether certain education levels are associated with higher capital gain and if there are gender-specific patterns within these categories.It aids in understanding the distribution of capital gain, helping in the identification of potential trends that can inform targeted marketing strategies.The visual supports the Gender-Specific Program Promotion user story, offering a nuanced view of the relationship between education, capital gain, and gender. Marketers can use these insights to tailor promotional initiatives based on the observed patterns in capital gain within different education levels and genders.

**9. Regional Enrolments Campaigns:**

The pair plot is designed to visualize the relationships between education, capital gain, and capital loss, with each data point colored by education level. The plot consists of scatter plots for pairs of variables and histograms along the diagonal. Education was chosen as the hue to distinguish different education levels, while capital gain and capital loss were selected as variables of interest.



**Conclusion:**

This pair plot aids in understanding how education level influences the relationships between capital gain and capital loss. Each scatter plot shows the distribution and correlation between two variables for different education levels. This visualization can assist in identifying patterns and trends, providing valuable insights for developing targeted enrolment campaigns based on education, capital gain, and capital loss, ultimately contributing to the goal of optimizing advertising budgets and increasing enrolment from diverse geographical areas.

**Questions:**

Below are the questions that arose during the progression of the project

***Feature Selection and Engineering:***

*Question:* Which features have the most significant impact on predicting income levels?

*Solution:* Conducted feature importance analysis using techniques like correlation analysis, feature importance scores from machine learning models and domain knowledge. Selected features based on their relevance to the prediction task and removed redundant and irrelevant features.

***Data Cleaning and Preprocessing:***

*Question*: How should missing values and inconsistent data be handled?

*Solution*: Implemented data cleaning techniques such as imputation for missing values, standardizing and encoding categorical variables. Used domain knowledge and exploratory data analysis to guide preprocessing decisions.

***Visualizations and Interpretation:***

*Question*: Which visualizations are most effective for understanding income distribution and demographic patterns?

*Solution*: Used a variety of visualizations such as histograms, bar plots, box plots, scatter plots, and heatmaps to explore relationships between variables and income levels. Interpreted visualizations to gain insights into demographic disparities and identify actionable patterns.

***Model Selection and Evaluation:***

*Question*: Which machine learning algorithms are suitable for predicting income levels?

*Solution*: Experimented with various algorithms such as Xgboost, Ensemble, Compose, and Imputer. Evaluate models using appropriate metrics like accuracy, precision, recall, F1-score, and ROC curves to select the best-performing model.

***Bias and Fairness Considerations:***

*Question*: How can potential biases in the data and model predictions be addressed?

*Solution*: Employed techniques like fairness-aware machine learning, bias detection, and mitigation strategies. Ensured fairness across demographic groups by monitoring model performance metrics for different subgroups and adjusting algorithms and data accordingly.

**Not doing:**

*Hyperparameter Tuning*: While I'm training machine learning models, I'm not extensively tuning hyperparameters. In the future, I intend to perform grid search or randomized search for hyperparameter optimization to fine-tune model performance.

*Feature Engineering*: Although I'm considering the importance of features, I'm not yet conducting extensive feature engineering such as creating interaction terms, polynomial features, or domain-specific transformations. Incorporating advanced feature engineering techniques can enhance model interpretability and predictive power.

**Appendix:**

**EDA and Visualization:**

import pandas as pd

import requests

from io import StringIO

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Replace 'your\_link\_here' with the actual link containing your data

url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data'

# Make a GET request to the URL

response = requests.get(url)

# Check if the request was successful (status code 200)

if response.status\_code == 200:

# Read the content of the response into a StringIO object

data = StringIO(response.text)

# Define column names based on your data

columns = ["age", "workclass", "fnlwgt", "education", "education\_num", "marital\_status",

"occupation", "relationship", "race", "sex", "capital\_gain", "capital\_loss",

"hours\_per\_week", "native\_country", "income"]

df = pd.read\_csv(data, names=columns, delimiter=', ', engine='python')

# Display the first few rows of the DataFrame

print(df.head())

else:

print(f"Failed to retrieve data. Status code: {response.status\_code}")

df.head(10)

missing\_values = df.isnull().sum()

print(missing\_values)

# Replace "?" with NaN

df.replace("?", pd.NA, inplace=True)

# Remove rows with NaN values

df = df.dropna()

df.duplicated().sum()

df.drop\_duplicates(inplace=True)

df.info()

df.describe()

"""Visual

### Business Analyst Stories:

1. Income Predictor

- Explore the correlation between education level and marital status.

- Determine the impact of the combination on an individual's income.

"""

def income\_category\_to\_numeric(category):

return 1 if category == '>50K' else 0

# function to create a new 'income\_numeric' column

df['income\_numeric'] = df['income'].apply(income\_category\_to\_numeric)

# Set up the visualization

plt.figure(figsize=(12, 8))

# Use Seaborn to create a heatmap

heatmap\_data = pd.pivot\_table(df, values='income\_numeric', index='education', columns='marital\_status')

sns.heatmap(heatmap\_data, annot=True, cmap="coolwarm", fmt=".2f", linewidths=.5)

# Set plot labels and title

plt.xlabel('Marital Status')

plt.ylabel('Education Level')

plt.title('Correlation between Education, Marital Status, and Income')

# Show the plot

plt.show()

"""2. Gender-Based Income Disparity

- Investigate the combination of gender and occupation.

- Identify potential gender-based income disparities.

"""

plt.figure(figsize=(14, 8))

sns.barplot(x='occupation', y='income\_numeric', hue='sex', data=df, ci=None, palette='pastel')

#labels and title

plt.xlabel('Occupation')

plt.ylabel('Average Income')

plt.title('Gender-Based Income Disparity by Occupation')

# legend

plt.legend(title='Gender', loc='upper right')

# Rotate x-axis labels

plt.xticks(rotation=45, ha='right')

plt.show()

"""3. Regional Income Variation

- Explore the interaction between native country and work class.

- Discern regional income variations for targeted strategies.

"""

# Set up the visualization

plt.figure(figsize=(16, 8))

sns.barplot(x='native\_country', y='income\_numeric', hue='workclass', data=df, errorbar=None, palette='viridis')

# plot labels and title

plt.xlabel('Native Country')

plt.ylabel('Average Income')

plt.title('Regional Income Variation by Work Class')

# legend

plt.legend(title='Work Class', loc='upper right')

plt.xticks(rotation=45, ha='right')

plt.show()

pivot\_table = pd.pivot\_table(df, values='income\_numeric', index='native\_country', columns='workclass', aggfunc='count', fill\_value=0)

# Find the country with the highest count for each work class

top\_countries\_by\_workclass = pivot\_table.idxmax(axis=0)

print("Top Contributing Countries by Work Class:")

print(top\_countries\_by\_workclass)

# Filter out entries from the United States

filtered\_df = df[df['native\_country'] != 'United-States']

# Create a pivot table for the filtered data

pivot\_table\_filtered = pd.pivot\_table(filtered\_df, values='income\_numeric', index='native\_country', columns='workclass', aggfunc='count', fill\_value=0)

# Find the country with the highest count for each work class in the filtered data

top\_countries\_by\_workclass\_filtered = pivot\_table\_filtered.idxmax(axis=0)

print("Top Contributing Countries by Work Class (excluding United-States):")

print(top\_countries\_by\_workclass\_filtered)

"""4.Diversity Impact:

- To analyze the correlation among race, occupation, and weekly working hours to gain insights into the impact on income diversity.

- Provide valuable information for promoting workplace inclusivity.

"""

plt.figure(figsize=(18, 8))

# Univariate analysis: Income distribution by race (box plot)

plt.subplot(1, 3, 1)

sns.boxplot(x='race', y='hours\_per\_week', hue='income', data=df, palette='bright', legend=False)

plt.title('Income Distribution by Race')

plt.xlabel('Race')

plt.ylabel('Hours Worked per Week')

plt.xticks(rotation=45, ha='right')

plt.show()

# Univariate analysis: Income distribution by occupation (box plot)

plt.figure(figsize=(38, 10))

plt.subplot(1, 3, 2)

sns.boxplot(x='occupation', y='hours\_per\_week', hue='income', data=df, palette='bright', legend=False)

plt.title('Income Distribution by Occupation')

plt.xlabel('Occupation')

plt.ylabel('Hours Worked per Week')

plt.xticks(rotation=45, ha='right')

plt.show()

# Univariate analysis: Income distribution by hours worked per week (box plot)

plt.figure(figsize=(20, 10))

plt.subplot(1, 3, 3)

sns.boxplot(x='income', y='hours\_per\_week', data=df, palette='bright')

plt.title('Income Distribution by Hours Worked per Week')

plt.xlabel('Income')

plt.ylabel('Hours Worked per Week')

plt.show()

"""### Marketing User Stories:

5. Targeted Marketing for Professions

- Leverage the combination of education, occupation, and work class.

- Tailor marketing campaigns for specific professions.

"""

plt.figure(figsize=(16, 8))

sns.barplot(x='education', y='income\_numeric', hue='occupation', data=df, errorbar=None, palette='muted')

# labels and title

plt.xlabel('Education Level')

plt.ylabel('Average Income')

plt.title('Targeted Marketing for Professions based on Education and Occupation')

#legend

plt.legend(title='Occupation', loc='upper right')

plt.xticks(rotation=45, ha='right')

plt.show()

"""6. Personalized Outreach for Marital Status

- Analyse the relationship between marital status, race, and hours worked per week.

- Create personalized outreach strategies for diverse audience segments.

"""

plt.figure(figsize=(18, 12))

plt.subplot(2, 2, 1)

sns.barplot(x='marital\_status', y='hours\_per\_week', hue='race', data=df,palette='pastel')

plt.title('Hours Worked per Week Distribution by Marital Status and Race')

plt.xticks(rotation=45, ha='right')

plt.subplot(2, 2, 2)

sns.lineplot(data=df, x='marital\_status', y='hours\_per\_week', hue='race', palette='pastel', marker='o', markersize=10, linestyle='-')

plt.title('Average Hours Worked per Week by Marital Status and Race')

plt.legend(loc='best', title='Race')

plt.tight\_layout()

plt.xticks(rotation=45, ha='right')

plt.show()

"""7. Cultural Tailoring

- Understand the impact of native country, race, and education on income.

- Tailor promotional content for diverse cultural backgrounds.

"""

plt.figure(figsize=(16, 8))

sns.boxplot(x='native\_country', y='income\_numeric', hue='race', data=df, palette='viridis')

# title and labels

plt.title('Income Distribution by Native Country and Race', fontsize=16)

plt.xlabel('Native Country')

plt.ylabel('Income')

plt.xticks(rotation=45, ha='right')

#legend

plt.legend(title='Race', bbox\_to\_anchor=(1.05, 0.8), loc='upper left')

plt.show()

"""8. Gender-Specific Program Promotion

- Investigate the combination of gender, education, and capital gain.

- Create gender-specific marketing initiatives.

"""

plt.figure(figsize=(16, 8))

sns.stripplot(x='education', y='capital\_gain', hue='sex', data=df, palette='coolwarm', size=5.5, jitter=True)

# title and labels

plt.title('Gender-Specific Program Promotion: Education, Capital Gain, and Gender', fontsize=16)

plt.xlabel('Education Level')

plt.ylabel('Capital Gain')

plt.xticks(rotation=45, ha='right')

# legend

plt.legend(title='Gender', bbox\_to\_anchor=(1.05, 0.8), loc='upper left')

plt.show()

"""

9. Regional Enrolments Campaigns

- Explore the relationship between native country, work class, and education.

- Develop targeted enrolments campaigns for specific regions.

"""

sns.pairplot(df, hue='education', vars=['capital\_gain', 'capital\_loss'], palette='Set2')

plt.suptitle('Pair Plot: Education, Capital Gain, and Capital Loss', y=1.02)

plt.show()

***Model:***

import pandas as pd

import numpy as np

import requests

import seaborn as sns

import matplotlib.pyplot as plt

import xgboost as xgb

from io import StringIO

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, classification\_report

from sklearn.preprocessing import OneHotEncoder

from sklearn.compose import ColumnTransformer

from sklearn.model\_selection import cross\_val\_score

from sklearn.metrics import confusion\_matrix

from sklearn.impute import SimpleImputer

url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data'

# Make a GET request to the URL

response = requests.get(url)

# Check if the request was successful (status code 200)

if response.status\_code == 200:

data = StringIO(response.text)

# Define column names based on your data

columns = ["age", "workclass", "fnlwgt", "education", "education\_num", "marital\_status",

"occupation", "relationship", "race", "sex", "capital\_gain", "capital\_loss",

"hours\_per\_week", "native\_country", "income"]

# Create a DataFrame from the CSV data

df = pd.read\_csv(data, names=columns, delimiter=', ', engine='python')

# Display the first few rows of the DataFrame

print(df.head())

else:

print(f"Failed to retrieve data. Status code: {response.status\_code}")

# Replace "?" with NaN

df.replace("?", pd.NA, inplace=True)

# Remove rows with NaN values

df = df.dropna()

missing\_values = df.isnull().sum()

print(missing\_values)

df.duplicated().sum()

df.drop\_duplicates(inplace=True)

df.info()

df.describe()

df.head()

import xgboost as xgb

from sklearn.model\_selection import cross\_val\_score

df['income\_numeric'] = df['income'].map({'<=50K': 0, '>50K': 1})

#'income' is our target variable

X = df.drop('income\_numeric', axis=1)

y = df['income\_numeric']

# categorical columns for one-hot encoding

categorical\_cols = X.select\_dtypes(include=['object']).columns.tolist()

# column transformer for preprocessing

preprocessor = ColumnTransformer(

transformers=[

('cat', OneHotEncoder(), categorical\_cols)

],

remainder='passthrough'

)

# preprocessing the feature matrix

X\_processed = preprocessor.fit\_transform(X)

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_processed, y, test\_size=0.2, random\_state=42)

# Instantiate XGBoost classifier

model = xgb.XGBClassifier()

# Train the model

model.fit(X\_train, y\_train)

# Make predictions on the test set

y\_pred = model.predict(X\_test)

# Evaluate accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

print("Model Accuracy:", accuracy)

# Evaluate the model

accuracy = accuracy\_score(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred)

# Print the results

print(f'Model Accuracy: {accuracy:.2f}')

cv\_scores = cross\_val\_score(model, X\_processed, y, cv=5)

print("Cross-Validation Scores:", cv\_scores)

print("Mean CV Accuracy:", cv\_scores.mean())

"""### 2. Identifying the most influential factors from the Census data to tailor marketing strategies effectively, ensuring enrolments increases while avoiding overreliance on specific demographics"""

# Select relevant features for the model

features = ['age', 'fnlwgt','workclass', 'education\_num', 'marital\_status', 'occupation', 'race', 'sex', 'capital\_gain', 'capital\_loss', 'hours\_per\_week', 'native\_country']

# Encode categorical features

df\_encoded = pd.get\_dummies(df[features])

X\_train, X\_test, y\_train, y\_test = train\_test\_split(df\_encoded, df['income\_numeric'], test\_size=0.2, random\_state=42)

# Instantiate RandomForestClassifier

rf\_model = RandomForestClassifier(random\_state=42)

# Train the model

rf\_model.fit(X\_train, y\_train)

# Get feature importances

feature\_importances = pd.DataFrame({'Feature': X\_train.columns, 'Importance': rf\_model.feature\_importances\_})

feature\_importances = feature\_importances.sort\_values(by='Importance', ascending=False)

# Display the top influential features

print("Top Influential Features:")

print(feature\_importances.head(10))

"""### 3. Addressing issues of missing or inconsistent data in the Census dataset, requiring robust data cleaning and imputation strategies for accurate and reliable model development."""

data = df.drop('income\_numeric', axis=1)

# Separate categorical and numerical columns

categorical\_cols = data.select\_dtypes(include=['object']).columns

numerical\_cols = data.select\_dtypes(exclude=['object']).columns

# Impute missing values with mode for categorical columns and median for numerical columns

imputer = SimpleImputer(strategy='most\_frequent', missing\_values=np.nan)

data\_imputed = data.copy()

data\_imputed[categorical\_cols] = imputer.fit\_transform(data[categorical\_cols])

imputer = SimpleImputer(strategy='median', missing\_values=np.nan)

data\_imputed[numerical\_cols] = imputer.fit\_transform(data[numerical\_cols])

# Verify that there are no missing values after imputation

print(data\_imputed.isnull().sum())

"""### 4. Bias Mitigation in Predictions: Tackling potential biases within the predictive model to ensure fair and unbiased outcomes, especially in sensitive areas such as income thresholds, to uphold ethical standards and inclusivity."""

# Evaluate the model with a focus on bias mitigation

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

print("Confusion Matrix:\n", conf\_matrix)

# Print classification report for detailed performance metrics

report = classification\_report(y\_test, y\_pred)

print("Classification Report:\n", report)

data\_imputed = data.copy()

for col in data\_imputed.columns:

if data\_imputed[col].dtype == 'object':

# Impute categorical columns with mode

data\_imputed[col].fillna(data\_imputed[col].mode()[0], inplace=True)

else:

# Impute numerical columns with median

data\_imputed[col].fillna(data\_imputed[col].median(), inplace=True)

print(data\_imputed.isnull().sum())

data\_imputed['income\_numeric'] = data\_imputed['income'].map({'<=50K': 0, '>50K': 1})

# Assuming 'income\_numeric' is your target variable

X = data\_imputed.drop(['income', 'income\_numeric'], axis=1)

y = data\_imputed['income\_numeric']

categorical\_cols = X.select\_dtypes(include=['object']).columns.tolist()

# Create a column transformer for preprocessing

preprocessor = ColumnTransformer(

transformers=[

('cat', OneHotEncoder(), categorical\_cols)

],

remainder='passthrough'

)

X\_processed = preprocessor.fit\_transform(X)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_processed, y, test\_size=0.2, random\_state=42)

# Instantiate XGBoost classifier

model = xgb.XGBClassifier()

model.fit(X\_train, y\_train)

# Make predictions on the test set

y\_pred = model.predict(X\_test)

# Evaluate accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

print("Model Accuracy:", accuracy)

# Evaluate the model with a focus on bias mitigation

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

print("Confusion Matrix:\n", conf\_matrix)