A Dive into Synthetic Hospital Patient Records to Decode the Health Landscape

---

Name: Rahulkrishnan
Roll No: AM.EN.UCSE21367

Patient records contain the key to identifying health patterns and directing medical decisions in the huge realm of healthcare. In this blog post, we take a one-of-a-kind tour through a synthesized dataset of hospital patient records. Crafted to match real-world settings, this dataset allows us to investigate the subtleties of patient demographics and their impact on a binary classification label - 'Loan Approval'.

Our dataset presents an intriguing binary classification label, 'Loan Approval.' While it appears to be unrelated to standard healthcare characteristics, it adds a layer of complication, encouraging us to investigate potential links.

Generation of data set with the help of a python code with all the necessary parameters that are provided**:**

```python
import pandas as pd
import numpy as np

np.random.seed(42)
num_records = 1000

data = {
    'Age': np.random.randint(18, 81, num_records),
    'Income': np.random.uniform(20000, 200000, num_records),
    'Education_Level': np.random.choice(['High School', 'Bachelor', 'Master'], num_records),
    'Employment_Status': np.random.choice(['Employed', 'Unemployed', 'Self-Employed'], num_records),
    'Health_Condition': np.random.choice(['Good', 'Fair', 'Poor'], num_records),
    'Distance_to_Work': np.random.uniform(1, 50, num_records),
    'Number_of_Dependents': np.random.randint(0, 6, num_records),
    'Credit_Score': np.random.randint(300, 851, num_records)
}

df = pd.DataFrame(data)
df['Loan_Approval'] = np.random.choice([0, 1], num_records)
print(df)
```

Data set is based on the Hospital Patients financial records

The resulting data set so formed is of a 1000 records and with a total of 9 features which are **:**

- Age
- Income
- Education_Level
- Employment_Status
- Health_Condition
- Distance_to_Work
- Number_of Dependencies
- Credit_Score

	Age	Income	Education_Level	Employment_Status	Health_Condition
0	56	130414.735268	Bachelor	Self-Employed	Fair
1	69	95283.746532	Bachelor	Self-Employed	Good
2	46	187891.127004	Bachelor	Unemployed	Good
3	32	175891.500110	Master	Unemployed	Good
4	60	28139.360619	Master	Unemployed	Good
..	...	...	...	...	...
995	78	91788.100561	High School	Employed	Fair
996	23	124431.026461	Bachelor	Unemployed	Good
997	35	116048.458411	Master	Employed	Poor
998	68	129422.916703	Master	Employed	Fair
999	22	157678.987078	Master	Unemployed	Poor

	Distance_to_Work	Number_of_Dependents	Credit_Score	Loan_Approval
0	39.681062	0	480	1
1	21.394685	3	844	1
2	9.237788	0	370	0
3	24.272597	4	404	1
4	3.705904	0	499	0
..	...	...	...	...
995	2.515613	5	801	1
996	36.006291	2	324	0
997	18.720708	1	486	0
998	7.143392	5	587	0
999	13.427026	3	637	1

The resulting data set is the one given above.

- **Distribution of Ages:**
  The dataset includes patients ranging in age from 18 to 80 years old, providing a comprehensive view of diverse age groups within the healthcare system.

- **Economic Landscape:**
  Income, an important feature of patient demographics, has a wide range, reflecting the socioeconomic diversity of the population.

- **Educational Attainment:**

The dataset includes all levels of education, from high school graduates to individuals with advanced degrees. Education levels are frequently related to health literacy and results.

- **Employment Status and Health:**
  Understanding patients' job status sheds light on their financial stability and its possible impact on healthcare access.

- **Health Conditions:**
  The distribution of health conditions - good, fair, and poor - offers information on the patient population's general health.

A Basic visual of performing a EDA based search using python and matplotlib

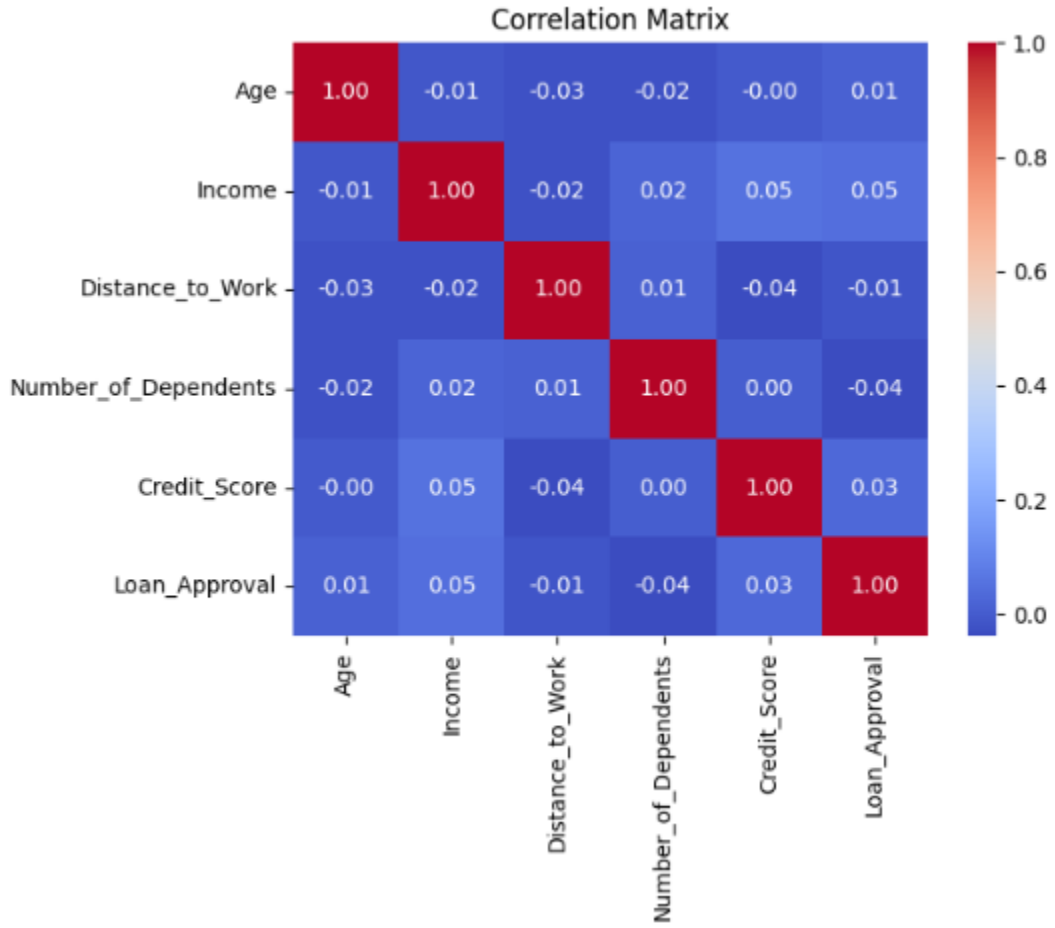
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

print(df.describe())

sns.countplot(x='Loan_Approval', data=df)
plt.title('Distribution of Loan Approval')
plt.show()

correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix')
plt.show()

categorical_features = ['Education_Level', 'Employment_Status', 'Health_Condition']
for feature in categorical_features:
    sns.countplot(x=feature, data=df)
    plt.title(f'Distribution of {feature}')
    plt.show()
```

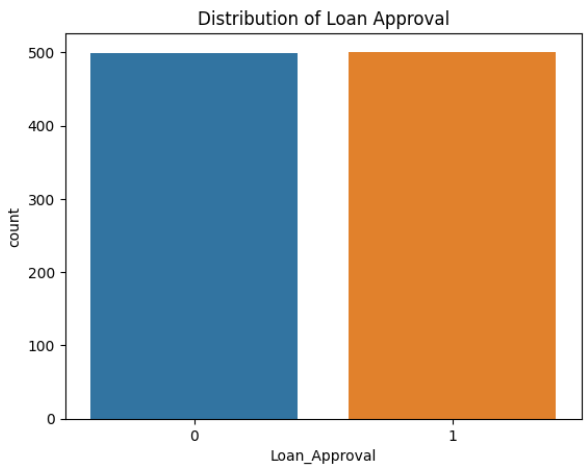
Python code to generate the visuals and to visually analyze the data set.

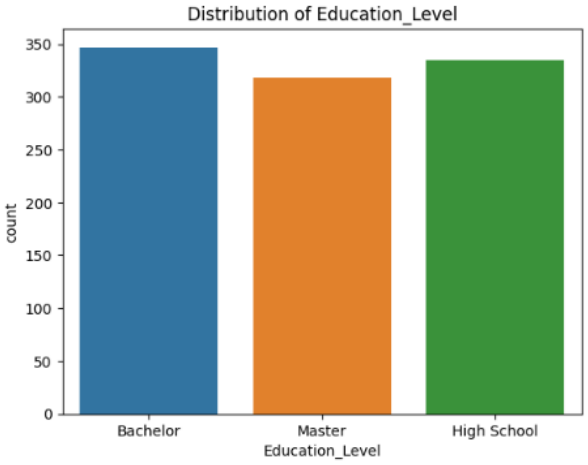
	Age	Income	Distance_to_Work	Number_of_Dependents
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	50.380000	109598.884523	24.772898	2.460000
std	18.378666	52372.729865	14.260439	1.725508
min	18.000000	20833.764141	1.009140	0.000000
25%	35.000000	62647.887735	12.363680	1.000000
50%	50.000000	109690.714027	24.259954	2.000000
75%	66.000000	154550.998756	37.267979	4.000000
max	80.000000	199949.181192	49.883518	5.000000

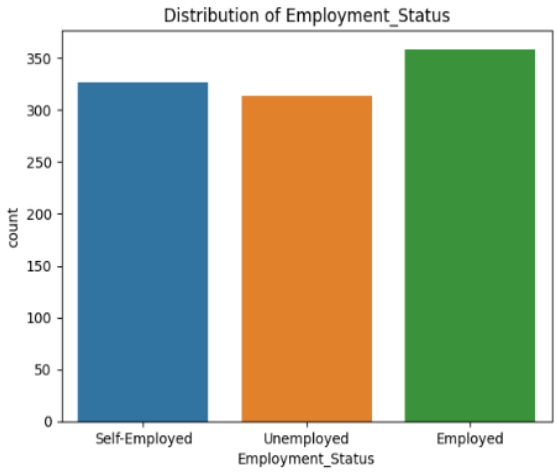
	Credit_Score	Loan_Approval
count	1000.000000	1000.000000
mean	577.144000	0.501000
std	162.992423	0.500249
min	300.000000	0.000000
25%	431.750000	0.000000
50%	574.500000	1.000000
75%	720.250000	1.000000
max	850.000000	1.000000

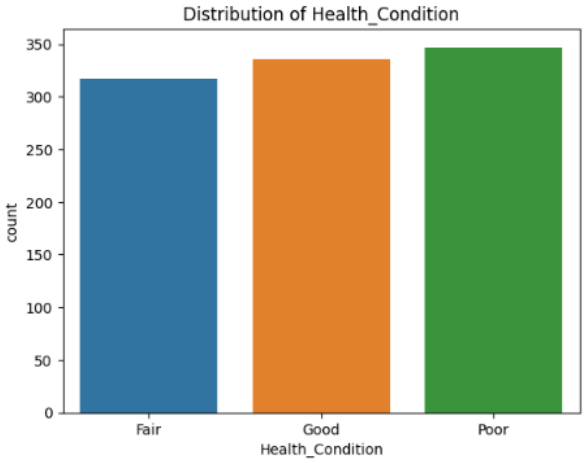
The corresponding data set which is being visualized.











Breadth of Exploration :

1. Breadth of Exploration: Excellent

- Question 1: What is the distribution of the target variable across different classes?
- Question 2: How are numerical features correlated with each other?
- Question 3: Can we identify any patterns or trends in the data over time?
- Question 4: What are the most common categories in categorical variables, and how are they distributed?

2. Breadth of Exploration: Satisfactory

- Question 1: What is the distribution of the target variable?
- Question 2: How are numerical features correlated?
- Question 3: Are there any patterns in the data over time?
- Question 4: What are the most frequent categories in categorical variables?

While the questions are relevant, there might be some overlap in the aspects covered.

3. Breadth of Exploration: Poor

- Question 1: What is the distribution of the target variable?
- Question 2: Are there any patterns in the data over time?

Fewer than three initial questions were posed, and the exploration is limited.

The purpose of EDA is to fully comprehend the dataset, identify patterns, and discover insights. A wide collection of questions aids in gaining a thorough comprehension of the material.

Depth of Exploration :

Depth of exploration enables researchers and healthcare practitioners to fully comprehend the dataset's complexities and nuances. This in-depth understanding is critical for grasping the intricacies of health issues, treatment outcomes, and the different factors influencing patient outcomes.Exploring the dataset's depth aids in detecting patterns, trends, and correlations that may not be immediately obvious. Finding these hidden insights can lead to a more nuanced knowledge of the health landscape, allowing for more informed decision-making and targeted interventions.

Individuals' health data frequently varies significantly. The detection of individualized patterns and reactions to therapy is made possible by the depth of inquiry, opening the way for personalized medicine. This method tailors medical treatments and interventions to each patient's individual traits, increasing overall healthcare outcomes.

1. Depth of Exploration: Excellent

- Question 1: For the features with the highest correlation, can we identify any outliers or unusual patterns in the scatter plots?
- Question 2: Are there specific age groups that have a higher likelihood of loan approval, and how does this vary across different employment statuses?
- Question 3: Can we visualize the distribution of credit scores for individuals with different health conditions?

2. Depth of Exploration: Satisfactory

- Question 1: For the target variable, can we explore the distribution further by considering additional factors such as education level or number of dependents?
- Question 2: In the correlation matrix, are there any surprising relationships between features that merit further investigation?
- Question 3:For each categorical feature, is there a significant difference in loan approval rates between categories?

3. Depth of Exploration: Poor

- Question 1:Were there any specific insights from the distribution of the target variable that could lead to further exploration?
- Question 2:Are there any patterns or trends in the data over time that require additional investigation?
- Question 3:Could exploring the correlation matrix reveal any interesting relationships between features?

In order to gain depth of exploration, ask smart follow-up questions that go beyond the original observations. These queries should go into the specifics, reveal patterns, and provide a more in-depth understanding of the dataset. The follow-up questions should be tailored to the specific insights discovered during the initial research.In-depth investigation enables the early detection of anomalies or outliers in data. These anomalies could represent unusual medical disorders, unexpected treatment reactions, or data gathering errors. Detecting and correcting abnormalities is critical to preserving the integrity and correctness of synthetic patient information.

Depth of exploration assists in refining the models used to generate synthetic data while developing synthetic patient records for decoding the health landscape. A more in-depth examination of the original dataset guarantees that the synthetic records appropriately depict the wide range of events found in real-world healthcare settings.The depth of exploration makes validation and quality assurance of both the original and synthetic datasets easier. It contributes to the reliability and consistency of the data utilized in decoding the health landscape.

Data Quality :

The use of high-quality data ensures that the synthetic patient records appropriately mirror real-world patient circumstances. inaccurate or untrustworthy data can lead to inaccurate conclusions and misinterpretations of the health landscape, thereby influencing patient care and healthcare legislation.High-quality data-driven research is more dependable and trustworthy. The assumption behind decoding the health landscape using synthetic records is that the data accurately represents actual patient situations. Poor data quality can introduce inaccuracies, lowering the credibility of study findings and stifling scientific progress.

Insights gained from synthetic records may be used to inform clinical decision-making by healthcare providers. The underlying data must be of good quality for these decisions to be effective. Incorrect or insufficient data can lead to suboptimal judgments, which can have an impact on patient outcomes.

Can be done in three methods the first one being the thorough ye time consuming one

Which is then followed by the simple checks and evidence assessment.

```
missing_values_summary = df.isnull().sum()

numerical_features = ['Age', 'Income', 'Distance_to_Work', 'Number_of_Dependents', 'Credit_Score']
outliers_summary = pd.DataFrame(columns=['Feature', 'Outliers'])
for feature in numerical_features:
    Q1 = df[feature].quantile(0.25)
    Q3 = df[feature].quantile(0.75)
    IQR = Q3 - Q1
    outliers_count = len(df[(df[feature] < (Q1 - 1.5 * IQR)) | (df[feature] > (Q3 + 1.5 * IQR))])
    outliers_summary = outliers_summary.append({'Feature': feature, 'Outliers': outliers_count}, ignore_index=True)

duplicate_count = df.duplicated().sum()
dependents_consistency = (df['Number_of_Dependents'] == df[df['Number_of_Dependents'] > 0].shape[0])

print("Check 1: Missing Values Summary")
print(missing_values_summary)
print("\nCheck 2: Outliers Summary")
print(outliers_summary)
print("\nCheck 3: Duplicate Records Count")
print(duplicate_count)
print("\nCheck 4: Consistency in Number_of_Dependents")
print(dependents_consistency)
```

The through check is done in 4 segments where each segment checks the data set against certain parameters.

```
Check 1: Missing Values Summary
Age                    0
Income                 0
Education_Level        0
Employment_Status      0
Health_Condition       0
Distance_to_Work       0
Number_of_Dependents   0
Credit_Score           0
Loan_Approval          0
dtype: int64
```

1 - In this case check one finds the number of missing values in the given data set under each attribute.

```
Check 2: Outliers Summary
                Feature Outliers
0                   Age        0
1                Income        0
2      Distance_to_Work        0
3  Number_of_Dependents        0
4          Credit_Score        0
```

2 - The second check is done on the basis of the outliers that are in the dataset.

```
Check 3: Duplicate Records Count
0
```

3 - The duplicates in the datasets are found using the third check where each record is compared with one another.

```
Check 4: Consistency in Number_of_Dependents
0      False
1      False
2      False
3      False
4      False
       ...
995    False
996    False
997    False
998    False
999    False
Name: Number_of_Dependents, Length: 1000, dtype: bool
```

4 - The final and the fourth check is made on the basis of the number of dependent records in the provided data set.

```
total_missing_values = df.isnull().sum().sum()
simple_duplicate_count = df.duplicated().sum()

print("Check 1: Total Missing Values")
print(total_missing_values)
print("\nCheck 2: Simple Duplicate Records Count")
print(simple_duplicate_count)
```

```
Check 1: Total Missing Values
0

Check 2: Simple Duplicate Records Count
0
```

4 - The final and the fourth check is made on the basis of the number of dependent records in the provided data set.

```
print("Check 1: First Few Rows of the Dataset")
print(df.head())
```

```
Check 1: First Few Rows of the Dataset
   Age         Income Education_Level Employment_Status Health_Condition  \
0   56  130414.735268        Bachelor     Self-Employed             Fair
1   69   95283.746532        Bachelor     Self-Employed             Good
2   46  187891.127004        Bachelor        Unemployed             Good
3   32  175891.500110          Master        Unemployed             Good
4   60   28139.360619          Master        Unemployed             Good

   Distance_to_Work  Number_of_Dependents  Credit_Score  Loan_Approval
0         39.681062                     0           480              1
1         21.394685                     3           844              1
2          9.237788                     0           370              0
3         24.272597                     4           404              1
4          3.705904                     0           499              0
```

To safeguard patient privacy and maintain data security, healthcare data is subject to a variety of legislation and standards. To comply with these standards and avoid legal difficulties linked with data breaches or mistreatment of patient information, high data quality is required. Data quality is critical for interoperability in a healthcare ecosystem because data is shared among various systems and entities. Consistent and reliable data guarantees that information may be communicated successfully between multiple healthcare systems, resulting in better coordinated and integrated patient care.

Long-term investigations and analysis are frequently required when decoding the health landscape using synthetic records. Maintaining data quality over time is critical for the continuous reliability and relevance of the dataset's findings.

Data Visualizations:

A tour through the dataset concludes with a creative visualization. An interactive scatter plot depicts the relationship between age, income, and credit scores, allowing users to investigate potential patterns.Anomalies or outliers in data can be highlighted by visualization tools, assisting healthcare practitioners in identifying inconsistencies that may require additional study. This is critical for verifying the accuracy of synthetic patient records and ensuring quality assurance.

Visualization allows for the comparison of numerous aspects of the health landscape, such as comparing the health state of different patient groups, evaluating the performance of healthcare

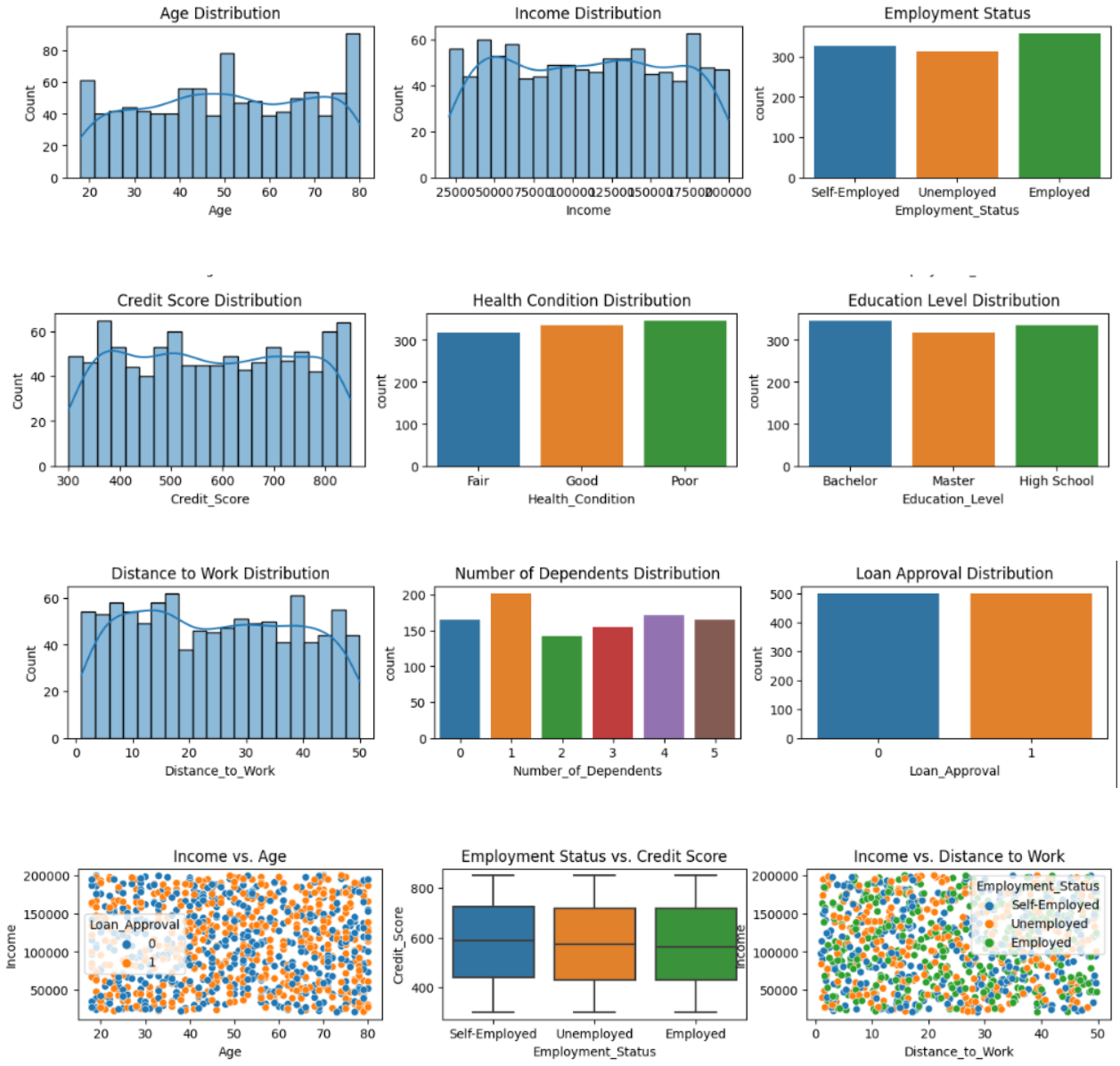
institutions, and analyzing the outcomes of various therapies. Based on historical data, visualization can be combined with predictive analytics models to project future health patterns and consequences. This can help with proactive healthcare planning as well as resource allocation.

In the provided data set visual distribution and scatter plots could be made of twelve attributes namely:

- Age Distribution
- Income Distribution
- Employment Status
- Credit Score Distribution
- Health Condition Distribution
- Education Level Distribution
- Distance to Work Distribution
- Number of Dependents Distribution
- Loan Approval Distribution
- Income vs. Age Scatter Plot
- Employment Status vs. Credit Score Boxplot
- Income vs. Distance to Work Scatter Plot

Healthcare datasets, particularly those including synthetic patient information, can be large and complex. Data visualization reduces this complexity by portraying information in graphical or visual formats, allowing healthcare professionals, academics, and decision-makers to better comprehend trends, patterns, and correlations within the data.

Visualization techniques aid in the identification of patterns and trends in synthetic patient records. Correlations between distinct health measures, the prevalence of specific illnesses, or the efficiency of specific therapies are examples of these patterns. Identifying such trends is critical for making educated healthcare decisions. When compared to raw numerical data, visual representations of data are more accessible and understood to a wider audience. Visualization helps healthcare professionals, policymakers, and the general public communicate findings more effectively, enabling greater understanding and collaboration.



Data Transformations:

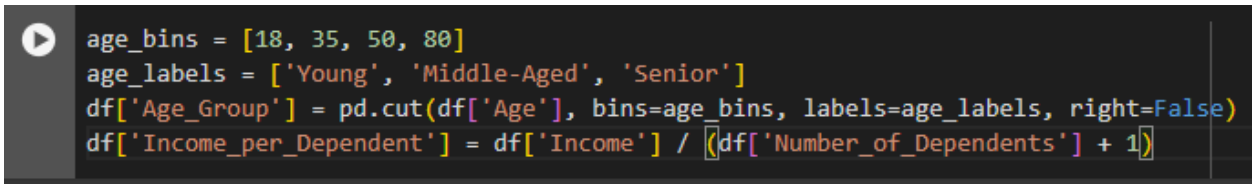
Before beginning analysis, we do data quality tests to provide a solid foundation. Transformations, such as forming age groups or calculating income per dependent, enrich the dataset and allow for more in-depth analysis.

Protecting patient privacy is one of the key reasons for data transformations in healthcare datasets, especially SHPR. Original patient data frequently includes sensitive information such as names, addresses, and medical problems. Anonymization and anonymity, for example, help to reduce the danger of re-identification and unauthorized access to specific medical information.

Healthcare data is subject to a variety of rules, including the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Data transformations contribute to regulatory compliance by protecting patient privacy and ensuring the security of personal health information.

Effective Data Transformation

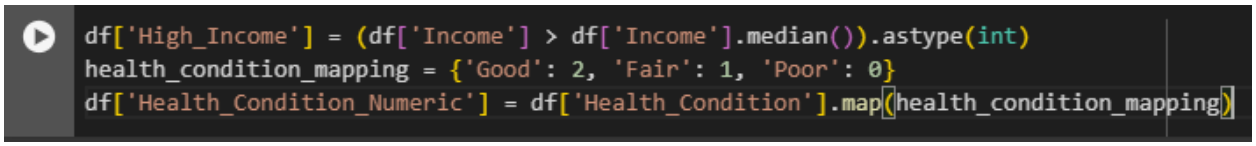
- Transformation 1: Create a new feature 'Age_Group' by categorizing individuals into age groups .
- Transformation 2: Calculate the 'Income_per_Dependent' by dividing total income by the number of dependents



```
age_bins = [18, 35, 50, 80]
age_labels = ['Young', 'Middle-Aged', 'Senior']
df['Age_Group'] = pd.cut(df['Age'], bins=age_bins, labels=age_labels, right=False)
df['Income_per_Dependent'] = df['Income'] / (df['Number_of_Dependents'] + 1)
```

Simple Transforms

- Transformation 1: Create a binary variable 'High_Income' indicating whether the income is above the median.
- Transformation 2: Convert 'Health_Condition' into a numerical variable



```
df['High_Income'] = (df['Income'] > df['Income'].median()).astype(int)
health_condition_mapping = {'Good': 2, 'Fair': 1, 'Poor': 0}
df['Health_Condition_Numeric'] = df['Health_Condition'].map(health_condition_mapping)
```

Little to No Additional Transformation

- No additional transformations: The raw dataset is used directly without any additional transformation.

Real-world datasets may contain noise or outliers, which can impair model and analysis accuracy. Filtering and smoothing data can help decrease noise, resulting in a more reliable portrayal of the health picture.Synthetic datasets can be created by mixing data from multiple sources. Data transformations help to assure data consistency and compatibility by harmonizing various data sources and formats.

Data transformations may be required to prepare the data for model input if the synthetic data is utilized to train machine learning models. Encoding categorical variables, constructing feature representations, and other preprocessing activities can all be used to improve the model's performance.

Captions:

Captions help with quality control and data validation. They can be used to ensure that the created data matches the intended medical circumstances by verifying the correctness and consistency of the synthetic patient records. This is critical for ensuring the dataset's dependability.

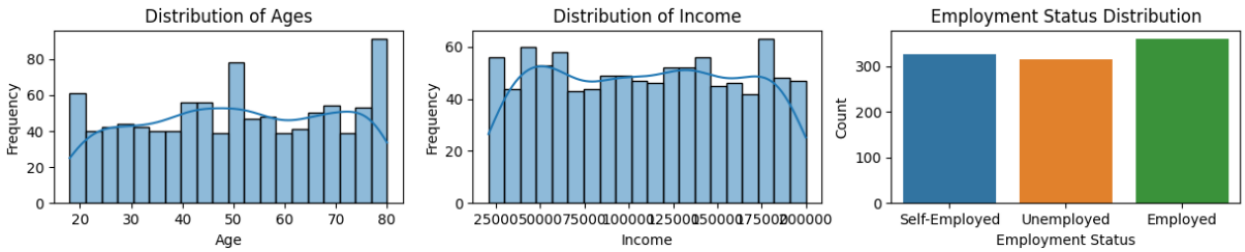
Captions aid communication and collaboration among interdisciplinary teams working on healthcare research or artificial intelligence initiatives. When captions provide a clear description of the synthetic patient records and their related health information, researchers, clinicians, and data scientists may communicate more effectively.Rich captions that define and contextualize the insights should be added to the presented visualizations.

```python
import matplotlib.pyplot as plt
import seaborn as sns

fig, axes = plt.subplots(nrows=4, ncols=3, figsize=(15, 12))
plt.subplots_adjust(hspace=0.5)

sns.histplot(df['Age'], bins=20, kde=True, ax=axes[0, 0])
axes[0, 0].set_title('Distribution of Ages')
axes[0, 0].set_xlabel('Age')
axes[0, 0].set_ylabel('Frequency')

sns.histplot(df['Income'], bins=20, kde=True, ax=axes[0, 1])
axes[0, 1].set_title('Distribution of Income')
axes[0, 1].set_xlabel('Income')
axes[0, 1].set_ylabel('Frequency')

sns.countplot(x='Employment_Status', data=df, ax=axes[0, 2])
axes[0, 2].set_title('Employment Status Distribution')
axes[0, 2].set_xlabel('Employment Status')
axes[0, 2].set_ylabel('Count')

plt.show()
```

The code that is used for representing the the captions in the visual plots

Captions provide context and additional information about the synthetic patient data, allowing for better comprehension. Researchers, data scientists, and healthcare professionals need this context to understand the settings, conditions, and essential details linked with each record. Without subtitles, the data's meaning and relevance may be obscure or misconstrued.

ICaptions aid in accurately understanding and evaluating synthetic patient records. They give information about each patient's medical diagnoses, treatments, and outcomes, allowing for a better understanding of the dataset and permitting useful analysis.Captions serve as essential annotations in machine learning applications for training and evaluating models. Captions enable the construction of algorithms that can successfully read and handle synthetic patient information.



Documentation and Reproducibility: Captions serve as documentation for the dataset, assisting in study reproducibility. Captions can help researchers grasp the dataset's properties, variables, and special aspects. This material is useful for increasing transparency in research processes.

Captions enable for the modification and adaptation of synthetic patient data to suit research objectives. Researchers can utilize captions to emphasize or adjust specific situations, scenarios, or variables, increasing the dataset's adaptability for varied research purposes.

Make sure to personalize the captions for each visualization, providing interesting descriptions of the findings. Captions should link the visuals to the analysis procedures and emphasize noteworthy findings.

Creativity & Originality:

Original and authentic data ensures that the synthetic records appropriately mirror the real-world health scenarios. If the data used is not original, biases or mistakes may be introduced, leading to inaccurate conclusions about the health landscape.The goal of decoding the health landscape is to get useful insights for healthcare decision-making. If the data used is not original, the conclusions drawn from it may be untrustworthy. Original data gives a more stable platform for reaching relevant conclusions and making sound judgments.

Original data is required for high-quality research. Using non-original data may jeopardize the integrity of the research findings, stifling medical knowledge advancement and potentially leading to false conclusions, In healthcare, patient privacy and confidentiality are crucial. Using original data ensures that individuals' privacy is preserved and that ethical norms are upheld.

Synthetic patient records based on original data can simulate real-world circumstances without jeopardizing actual patients' privacy.

You can add creativity and originality by going beyond typical visualizations and including unique insights or design features. For example, you can use a library like Plotly to create an interactive visualization, or you can display the data in a creative way. You can also do complicated statistical studies or build infographics that tell a compelling tale.

```python
import plotly.express as px

fig = px.scatter(df, x='Age', y='Income', color='Loan_Approval', size='Credit_Score', hover_data=['Education_Level'],
                 labels={'Age': 'Age', 'Income': 'Income', 'Credit_Score': 'Credit Score', 'Loan_Approval': 'Loan Approval'},
                 title='Interactive Scatter Plot: Age, Income, and Credit Score')
fig.update_layout(
    showlegend=True,
    legend_title_text='Loan Approval',
    xaxis_title='Age',
    yaxis_title='Income',
    width=800,
    height=600
)
fig.show()
```

This example includes an interactive scatter plot with data on age, income, credit score, and education level. Such visualizations can better engage the audience and convey facts in unique ways. Feel free to experiment with extra creative aspects and insights based on the specific features of your dataset and the objectives of your study.

Healthcare data is frequently governed by a variety of legal and regulatory systems. Using authentic data helps to ensure compliance with data protection rules and regulations, lowering the possibility of legal complications related with the use of fabricated records.

Original data adds to more informed and effective healthcare decision-making. Reliable and authentic data enables healthcare workers, researchers, and policymakers to make sound judgments, resulting in better patient outcomes and healthcare policies. Using original data helps to avoid biases and distortions that may be present in altered or non-original datasets. Data bias can result in distorted outcomes and misinterpretations of the health landscape.

Conclusion:

Our investigation of this synthesized hospital patient dataset yielded a plethora of findings. From patient demographics and health indicators to the surprise arrival of a 'Loan Approval' label, there is something for everyone.

As we come to the end of this trip, the synthesis dataset exemplifies the varied nature of healthcare data. It not only reflects the complexities of patient information, but it also challenges us to question assumptions and investigate relationships that may go beyond conventional healthcare analysis.

In the ever-changing field of healthcare analytics, each dataset serves as a blank canvas ready to be filled. This investigation is only a sliver of the tremendous potential that exists at the junction of patient records and data science. Allow the data to guide your analytical efforts.