task_1_eda

December 28, 2024

1 Exploratory Data Analysis (EDA) for Insurance Data Analysis

1.1 What this notebook does is:

- Load the data
- Perform basic statistics on the data

[1]:

```python
# Import necessary libraries
import sys
import os
import matplotlib.pyplot as plt
import pandas as pd
```

[2]:

```python
# Get the current working directory of the project
current_dir = os.getcwd()
print(current_dir)

# Get the parent directory
parent_dir = os.path.dirname(current_dir)
print(parent_dir)

# Insert the path to the parent directory
sys.path.insert(0, parent_dir)

# # Insert the path to the Scripts directory
# sys.path.insert(0, os.path.join(parent_dir, 'Scripts'))

# print(sys.path)
```

```
c:\Users\HP\Desktop\KAIM-Cohort-3\Week 3\AlphaCare-Insurance-Solutions-(ACIS)-Insurance-Claim-Data Analysis\notebooks
c:\Users\HP\Desktop\KAIM-Cohort-3\Week 3\AlphaCare-Insurance-Solutions-(ACIS)-Insurance-Claim-Data Analysis
```

[3]:

```python
## 1. Load Data
from scripts.eda_utils import load_data, summarize_data
from scripts.plot_utils import plot_histogram, plot_correlation_matrix, plot_boxplot
```

[4]:

```
# Load the dataset
file_path = "../data/MachineLearningRating_v3.txt"
data = pd.read_csv(file_path, delimiter="|")
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_21496\4107697220.py:3: DtypeWarning:
Columns (32,37) have mixed types. Specify dtype option on import or set
low_memory=False.
  data = pd.read_csv(file_path, delimiter="|")
```

[5]:

```
# Convert 'TransactionMonth' to datetime
if 'TransactionMonth' in data.columns:
    data['TransactionMonth'] = pd.to_datetime(data['TransactionMonth'],␣
 ↪errors='coerce')
```

[6]:

```
# Check for columns with non-numeric data
non_numeric_cols = data.select_dtypes(include=['object']).columns
print(f"Non-numeric columns: {non_numeric_cols}")
```

```
Non-numeric columns: Index(['Citizenship', 'LegalType', 'Title', 'Language',
'Bank', 'AccountType',
       'MaritalStatus', 'Gender', 'Country', 'Province', 'MainCrestaZone',
       'SubCrestaZone', 'ItemType', 'VehicleType', 'make', 'Model', 'bodytype',
       'VehicleIntroDate', 'AlarmImmobiliser', 'TrackingDevice',
       'CapitalOutstanding', 'NewVehicle', 'WrittenOff', 'Rebuilt',
       'Converted', 'CrossBorder', 'TermFrequency', 'ExcessSelected',
       'CoverCategory', 'CoverType', 'CoverGroup', 'Section', 'Product',
       'StatutoryClass', 'StatutoryRiskType'],
      dtype='object')
```

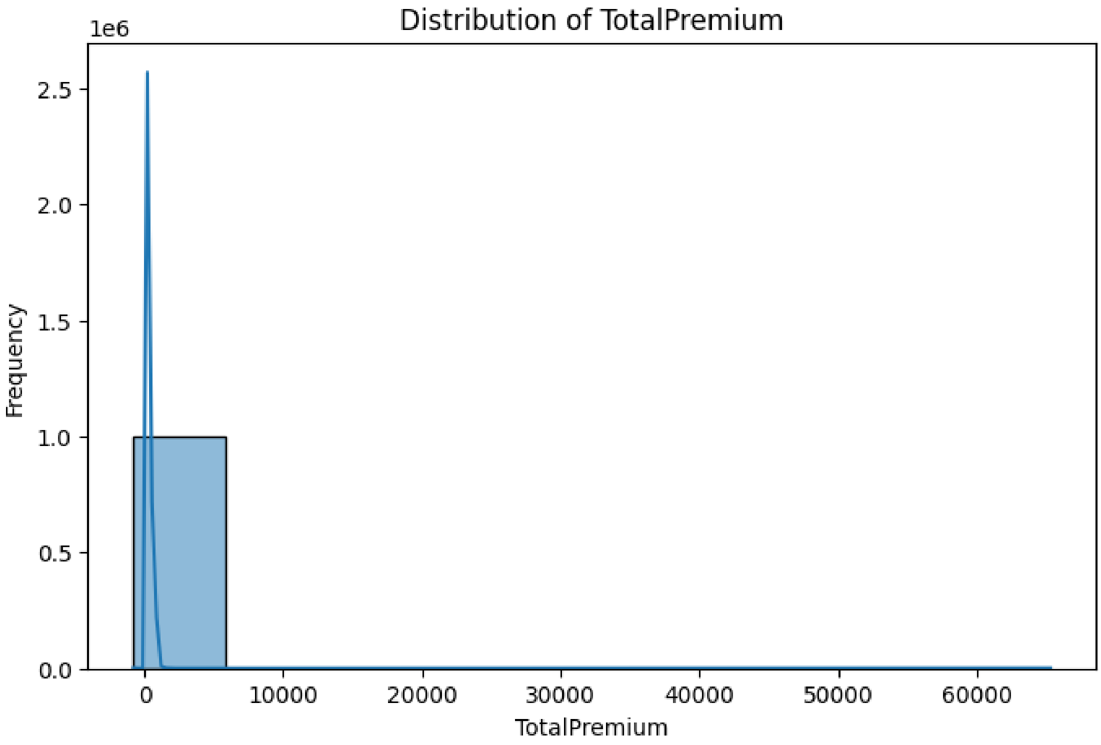
[7]:

```
# Drop or exclude non-numeric columns for numerical operations
numeric_data = data.select_dtypes(include=['number'])
print("Prepared data for numerical operations.")
```

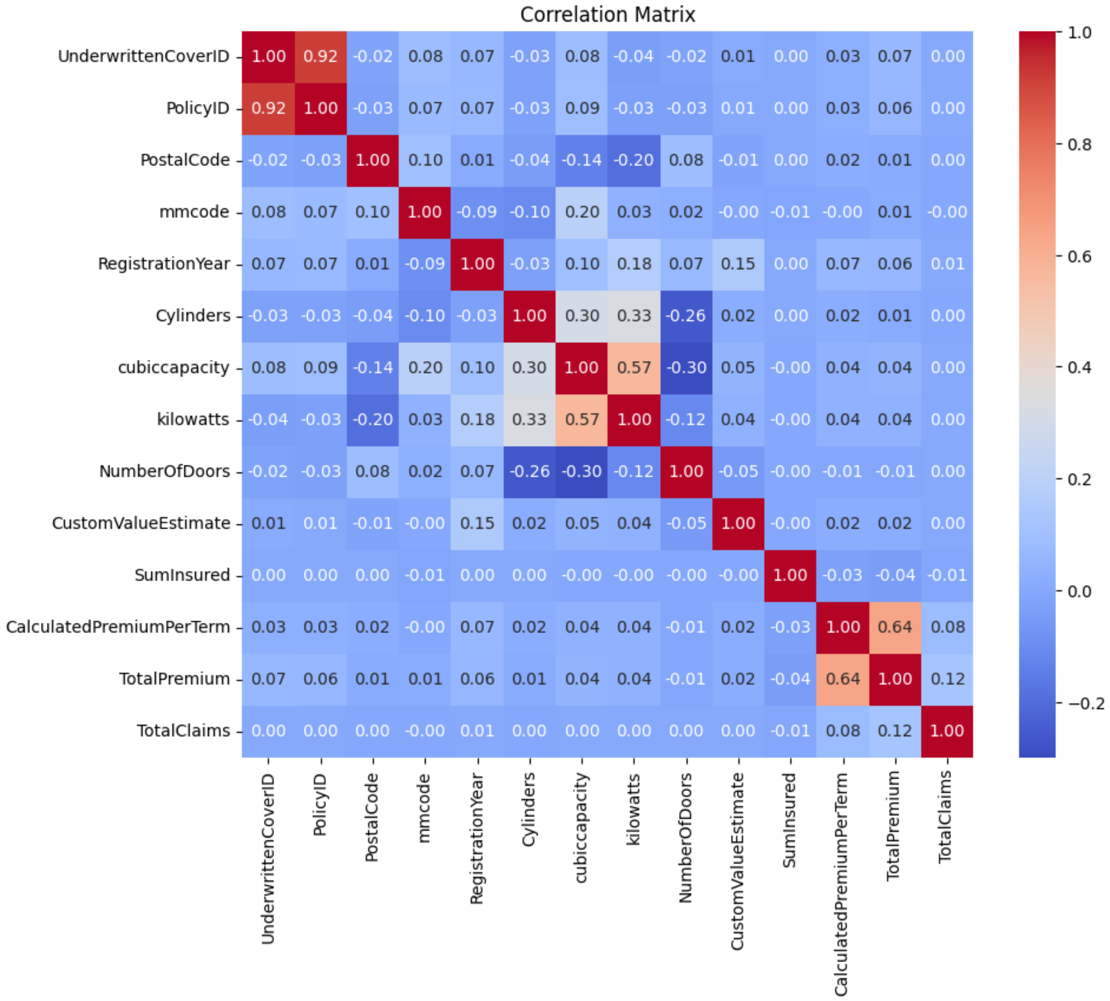
```
Prepared data for numerical operations.
```

[8]:

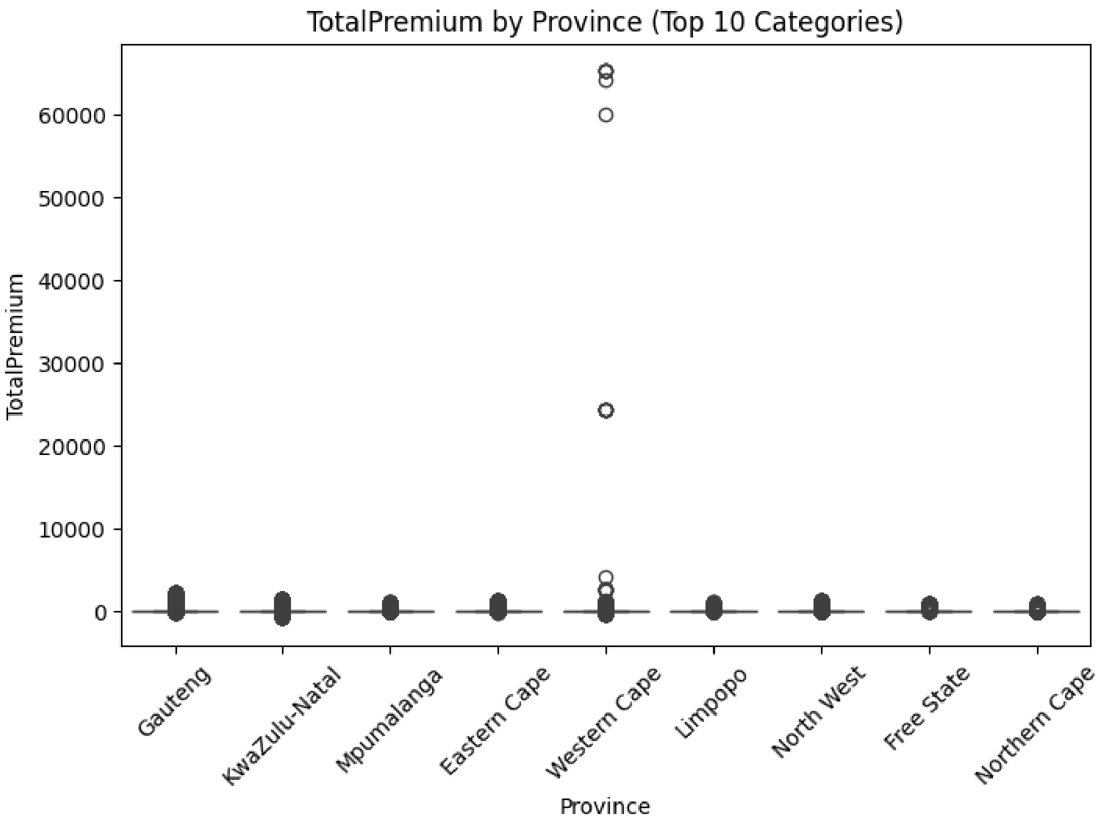
```
# Plot Histogram for TotalPremium
plot_histogram(data, "TotalPremium")
```



[9]:
```
# Plot Correlation Matrix
plot_correlation_matrix(data)
```



```
[10]: # Plot Boxplot for TotalPremium by Province
      plot_boxplot(data, "TotalPremium", "Province", max_categories=10)
```


TotalPremium by Province (Top 10 Categories)
TotalPremium
60000
50000
40000
30000
20000
10000
0
Gauteng
KwaZulu-Natal
Mpumalanga
Eastern Cape
Western Cape
Limpopo
North West
Free State
Northern Cape
Province