task_2_dvc

December 28, 2024

1 Data Version Control (DVC) Analysis of Insurance Data Notebook

[2]:

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

[3]:

```
# 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
```

[4]:

```
## Step 1: Initialize DVC
os.chdir(parent_dir)
from scripts.dvc_utils import init_dvc, add_remote, track_file, push_data
```

[5]:

```
remote_path = "../data/dvc_storage"
print(f"Adding remote storage at {remote_path}...")
add_remote(remote_path)
```

```
Adding remote storage at ../data/dvc_storage…
Adding remote storage at: ../data/dvc_storage
Setting 'localstorage' as a default remote.
```

[6]:

```python
## Step 2: Track the Dataset
file_path = "../data/MachineLearningRating_v3.txt"
print(f"Tracking dataset: {file_path}...")
track_file(file_path)
```

```
Tracking dataset: ../data/MachineLearningRating_v3.txt…
Tracking file: ../data/MachineLearningRating_v3.txt
Error occurred while tracking the file: ERROR: Cached output(s) outside of DVC
project: c:\Users\HP\Desktop\KAIM-Cohort-3\Week
3\data\MachineLearningRating_v3.txt. See <https://dvc.org/doc/user-guide/data-
management/importing-external-data> for more info.
```

[7]:

```python
## Step 3: Push Data to Remote Storage
print("Pushing data to remote storage...")
push_data()
```

```
Pushing data to remote storage…
Pushing data to remote storage…
Everything is up to date.
```

[8]:

```python
## Step 4: Visualization (Boxplot for Top 10 Categories)
from scripts.plot_utils import plot_boxplot
import pandas as pd
```

[9]:

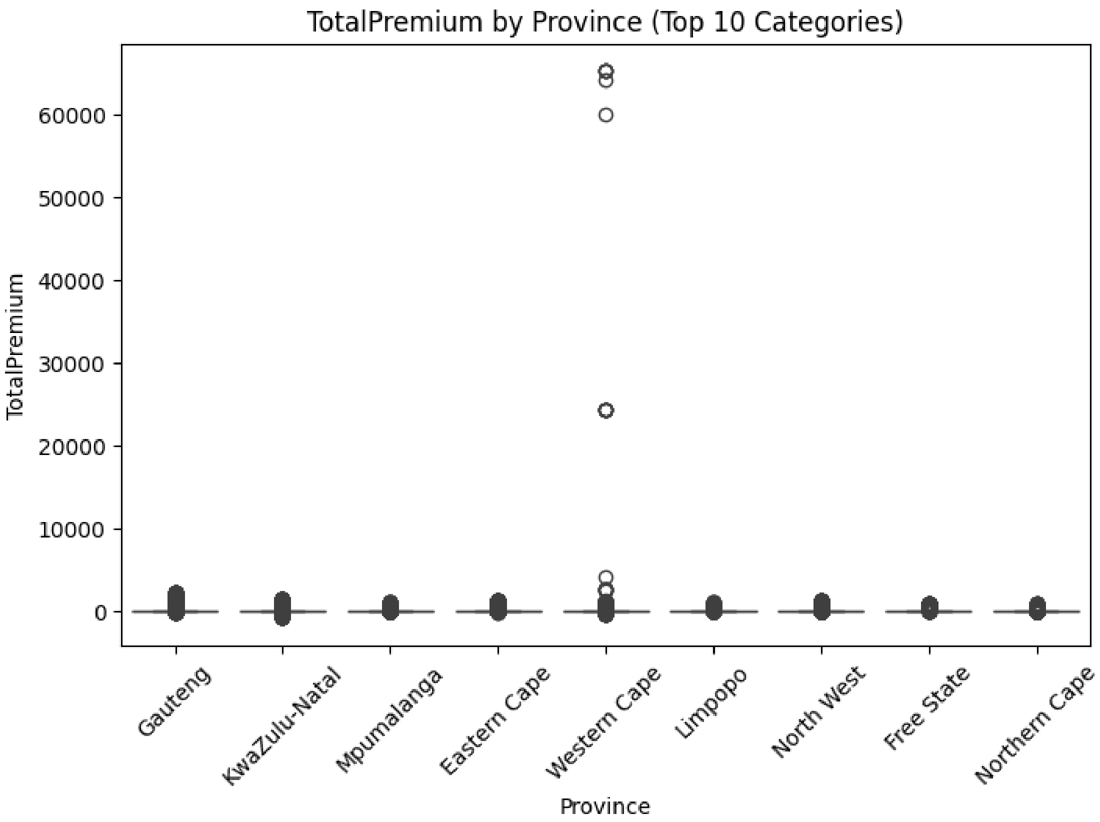
```python
# Load the dataset
data = pd.read_csv(file_path, delimiter="|")
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_18300\3844965664.py:2: DtypeWarning:
Columns (32,37) have mixed types. Specify dtype option on import or set
low_memory=False.
  data = pd.read_csv(file_path, delimiter="|")
```

[10]:

```python
# Plot boxplot for TotalPremium grouped by Province (only top 10 categories)
plot_boxplot(data, "TotalPremium", "Province", max_categories=10)

print("Task 2 completed successfully.")
```



```
Task 2 completed successfully.
```