02_eda

November 26, 2025

1 Day 6-7: Exploratory Data Analysis (EDA)

[89]:
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1. Basic Statistics

[90]:
```python
df = pd.read_csv('../data/processed/cleaned_data.csv')
print(df.describe())
print('-----------------------------------------------------------------------------------------
print(df['FraudFound_P'].value_counts())  # Check class balance
df
```

```
       WeekOfMonth  WeekOfMonthClaimed           Age  FraudFound_P  \
count  15419.000000        15419.000000  15419.000000  15419.000000
mean       2.788637            2.694079     40.685842      0.059861
std        1.287611            1.259082     12.181893      0.237237
min        1.000000            1.000000     16.000000      0.000000
25%        2.000000            2.000000     31.000000      0.000000
50%        3.000000            3.000000     39.000000      0.000000
75%        4.000000            4.000000     48.000000      0.000000
max        5.000000            5.000000     80.000000      1.000000

      PolicyNumber     RepNumber    Deductible  DriverRating          Year
count   15419.00000  15419.000000  15419.000000  15419.000000  15419.000000
mean     7710.90168      8.482846    407.704780      2.487840   1994.866528
std      4451.37980      4.599798     43.952379      1.119482      0.803309
min         1.00000      1.000000    300.000000      1.000000   1994.000000
25%      3856.50000      5.000000    400.000000      1.000000   1994.000000
50%      7711.00000      8.000000    400.000000      2.000000   1995.000000
75%     11565.50000     12.000000    400.000000      3.000000   1996.000000
max     15420.00000     16.000000    700.000000      4.000000   1996.000000
--------------------------------------------------------------------------------
---------
FraudFound_P
0    14496
1      923
```

```
Name: count, dtype: int64
```

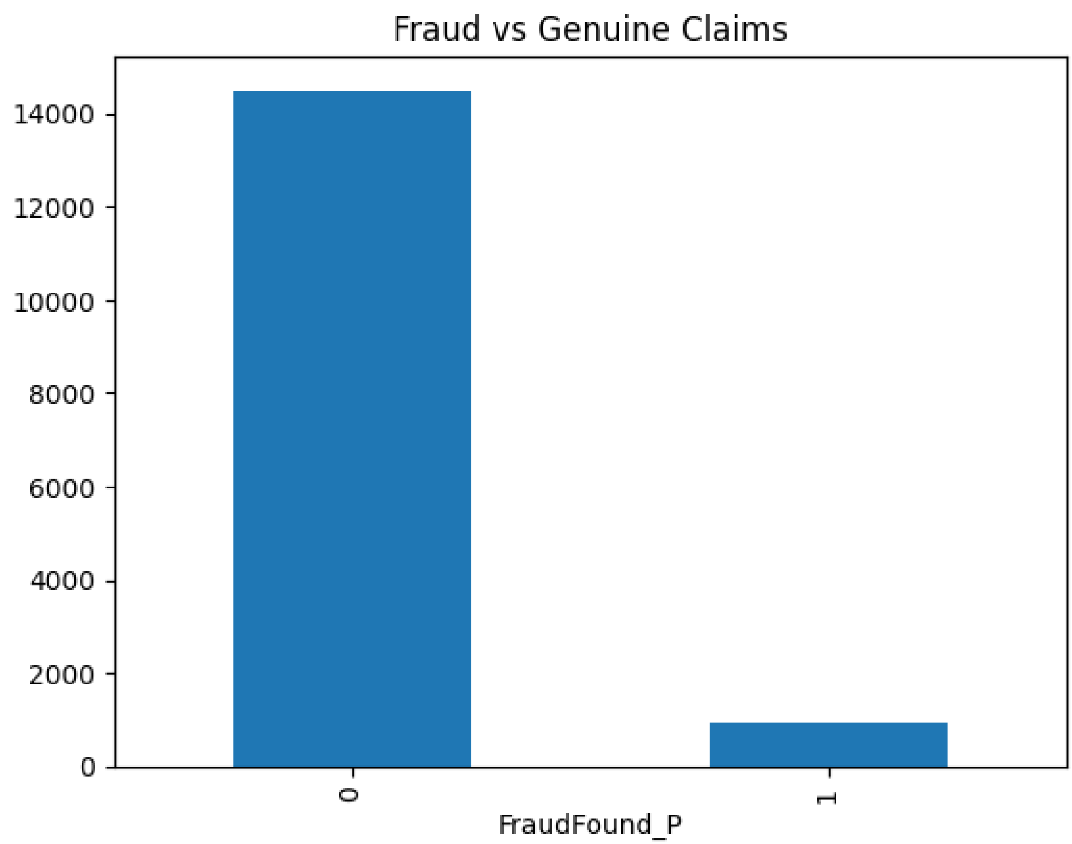
[90]:

```
      Month  WeekOfMonth  DayOfWeek    Make AccidentArea DayOfWeekClaimed  \
0       Dec            5  Wednesday   Honda        Urban          Tuesday
1       Jan            3  Wednesday   Honda        Urban           Monday
2       Oct            5     Friday   Honda        Urban         Thursday
3       Jun            2   Saturday  Toyota        Rural           Friday
4       Jan            5     Monday   Honda        Urban          Tuesday
…       …          …           …      …          …              …
15414   Nov            4     Friday  Toyota        Urban          Tuesday
15415   Nov            5   Thursday Pontiac        Urban           Friday
15416   Nov            5   Thursday  Toyota        Rural           Friday
15417   Dec            1     Monday  Toyota        Urban         Thursday
15418   Dec            2  Wednesday  Toyota        Urban         Thursday

      MonthClaimed  WeekOfMonthClaimed     Sex MaritalStatus  …  \
0              Jan                   1  Female        Single  …
1              Jan                   4    Male        Single  …
2              Nov                   2    Male       Married  …
3              Jul                   1    Male       Married  …
4              Feb                   2  Female        Single  …
…            …                  …       …            …  …
15414          Nov                   5    Male       Married  …
15415          Dec                   1    Male       Married  …
15416          Dec                   1    Male        Single  …
15417          Dec                   2  Female       Married  …
15418          Dec                   3    Male        Single  …

       AgeOfVehicle AgeOfPolicyHolder PoliceReportFiled WitnessPresent  \
0           3 years          26 to 30                No             No
1           6 years          31 to 35               Yes             No
2           7 years          41 to 50                No             No
3       more than 7          51 to 65               Yes             No
4           5 years          31 to 35                No             No
…               …                 …               …            …
15414       6 years          31 to 35                No             No
15415       6 years          31 to 35                No             No
15416       5 years          26 to 30                No             No
15417       2 years          31 to 35                No             No
15418       5 years          26 to 30                No             No

      AgentType  NumberOfSuppliments  AddressChange_Claim  NumberOfCars  Year  \
0      External                 none               1 year        3 to 4  1994
1      External                 none            no change     1 vehicle  1994
2      External                 none            no change     1 vehicle  1994
3      External          more than 5            no change     1 vehicle  1994
4      External                 none            no change     1 vehicle  1994
```

```
…        …                   …               …           …    …
15414  External            none         no change   1 vehicle  1996
15415  External     more than 5         no change      3 to 4  1996
15416  External          1 to 2         no change   1 vehicle  1996
15417  External     more than 5         no change   1 vehicle  1996
15418  External          1 to 2         no change   1 vehicle  1996

        BasePolicy
0        Liability
1        Collision
2        Collision
3        Liability
4        Collision
…            …
15414    Collision
15415    Liability
15416    Collision
15417  All Perils
15418    Collision

[15419 rows x 33 columns]
```

2. Fraud Distribution

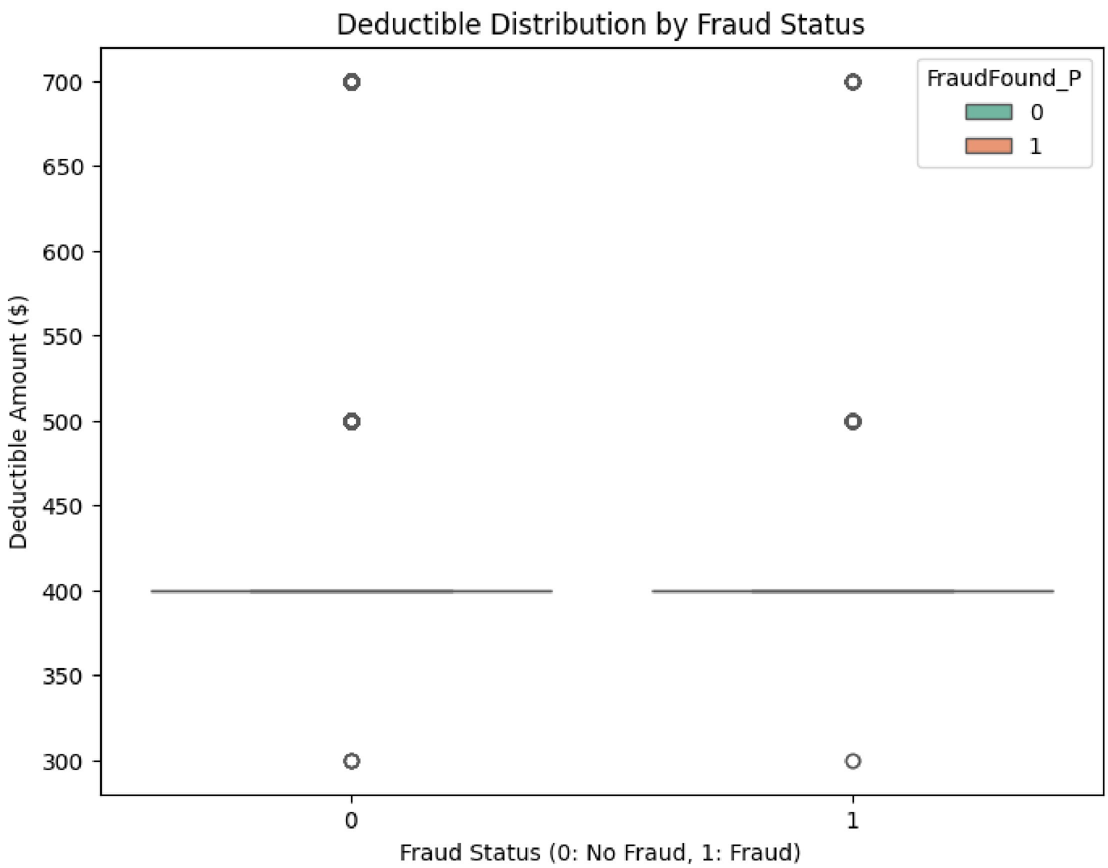
[91]:
```
df['FraudFound_P'].value_counts().plot(kind='bar')
plt.title('Fraud vs Genuine Claims')
plt.savefig('EDA_Charts/Fraud vs Genuine Claims.png')
plt.show()
```



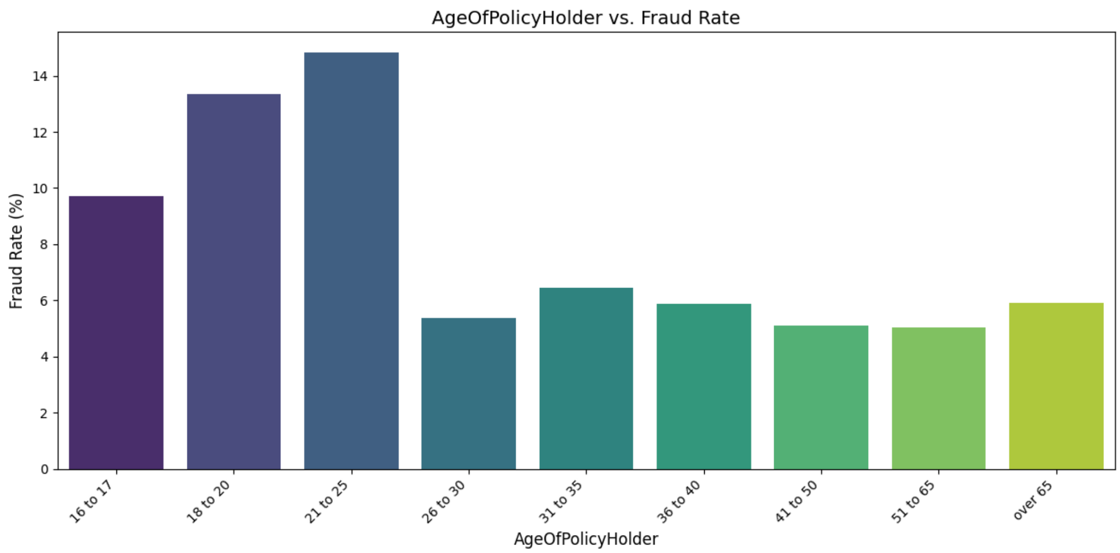
3. Analyze Key Features

[92]:
```
# 1. Age Distribution by Fraud
plt.figure(figsize=(10, 5))
df[df['FraudFound_P'] == 0]['Age'].hist(bins=20, alpha=0.5, label='Genuine␣
↪(0)', color='green')
df[df['FraudFound_P'] == 1]['Age'].hist(bins=20, alpha=0.5, label='Fraud (1)',␣
↪color='red')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age Distribution: Fraud vs Genuine')
plt.legend()
plt.savefig('EDA_Charts/age_distribution_fraud.png')
plt.show()
```

[93]:
```
# 2. Total Claim Amount (sum of all claim types)
plt.figure(figsize=(8, 6))
sns.boxplot(x='FraudFound_P', y='Deductible', data=df, palette='Set2',␣
 ↪hue='FraudFound_P')
plt.title('Deductible Distribution by Fraud Status')
plt.xlabel('Fraud Status (0: No Fraud, 1: Fraud)')
plt.ylabel('Deductible Amount ($)')
plt.savefig('deductible_boxplot.png')
plt.show()
```

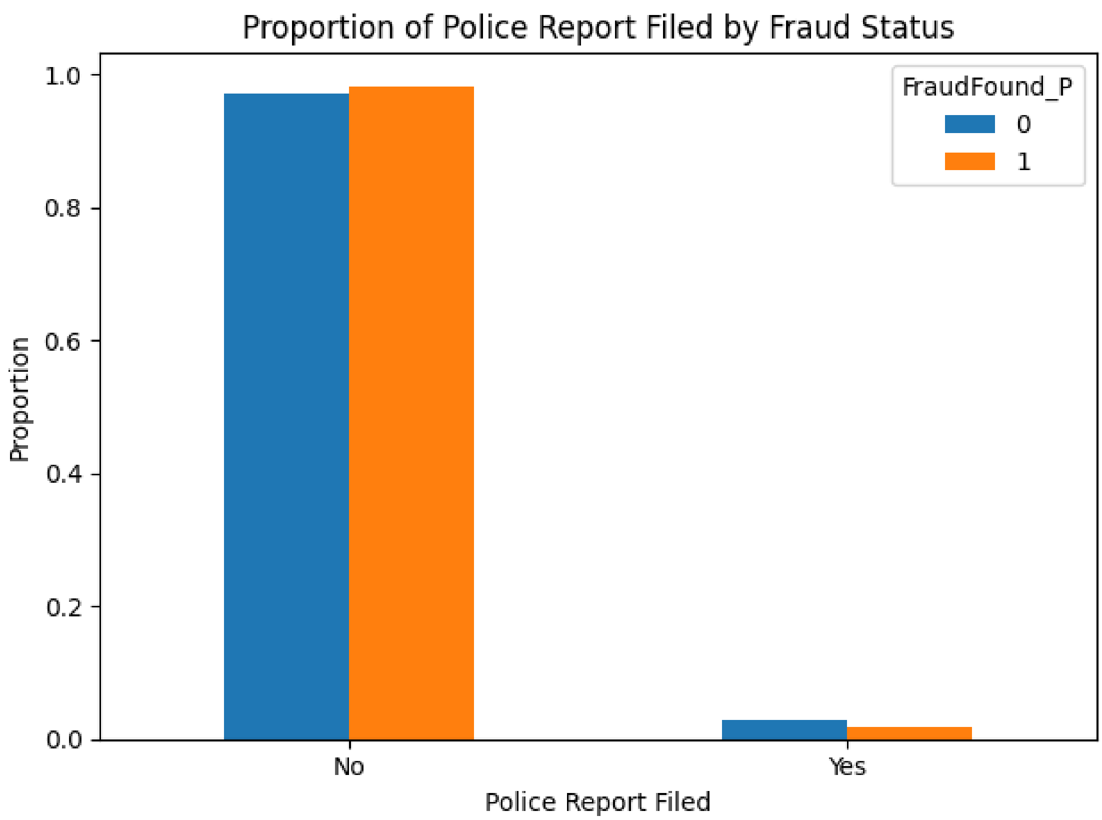


[94]:
```
plt.figure(figsize=(12, 6))
fraud_rate = df.groupby('AgeOfPolicyHolder')['FraudFound_P'].mean()
sns.barplot(    x=fraud_rate.index, y=fraud_rate.values * 100,␣
↪palette='viridis', hue=fraud_rate.index, legend=False)
plt.title(f'{'AgeOfPolicyHolder'} vs. Fraud Rate', fontsize=14)
plt.ylabel('Fraud Rate (%)', fontsize=12)
plt.xlabel('AgeOfPolicyHolder', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.savefig(f'EDA_Charts/{'AgeOfPolicyHolder'}_fraud_rate.png')
plt.show()
```

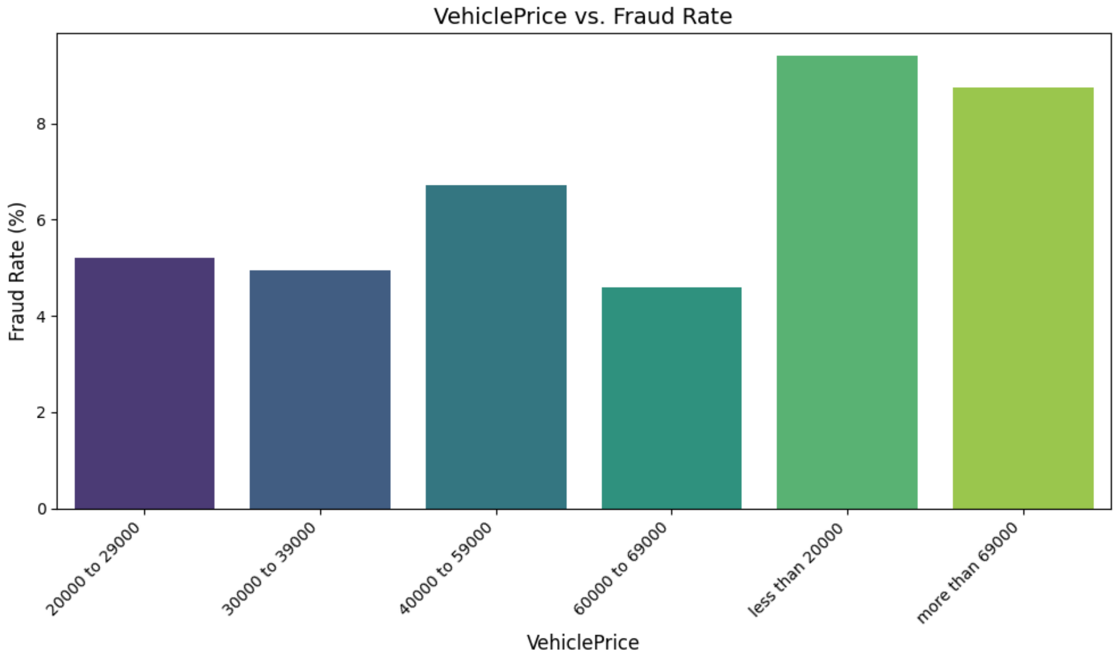


[95]:
```
# 3. Police Report Filed
plt.figure(figsize=(7, 5))
pd.crosstab(df['PoliceReportFiled'], df['FraudFound_P'], normalize='columns').
  ↪plot(kind='bar')
plt.title('Proportion of Police Report Filed by Fraud Status')
plt.xlabel('Police Report Filed')
plt.ylabel('Proportion')
plt.legend(title='FraudFound_P')
plt.xticks(rotation=0)
plt.tight_layout()
plt.savefig('EDA_Charts/police_report_filed_fraud_proportion.png')
plt.show()
```

```
<Figure size 700x500 with 0 Axes>
```



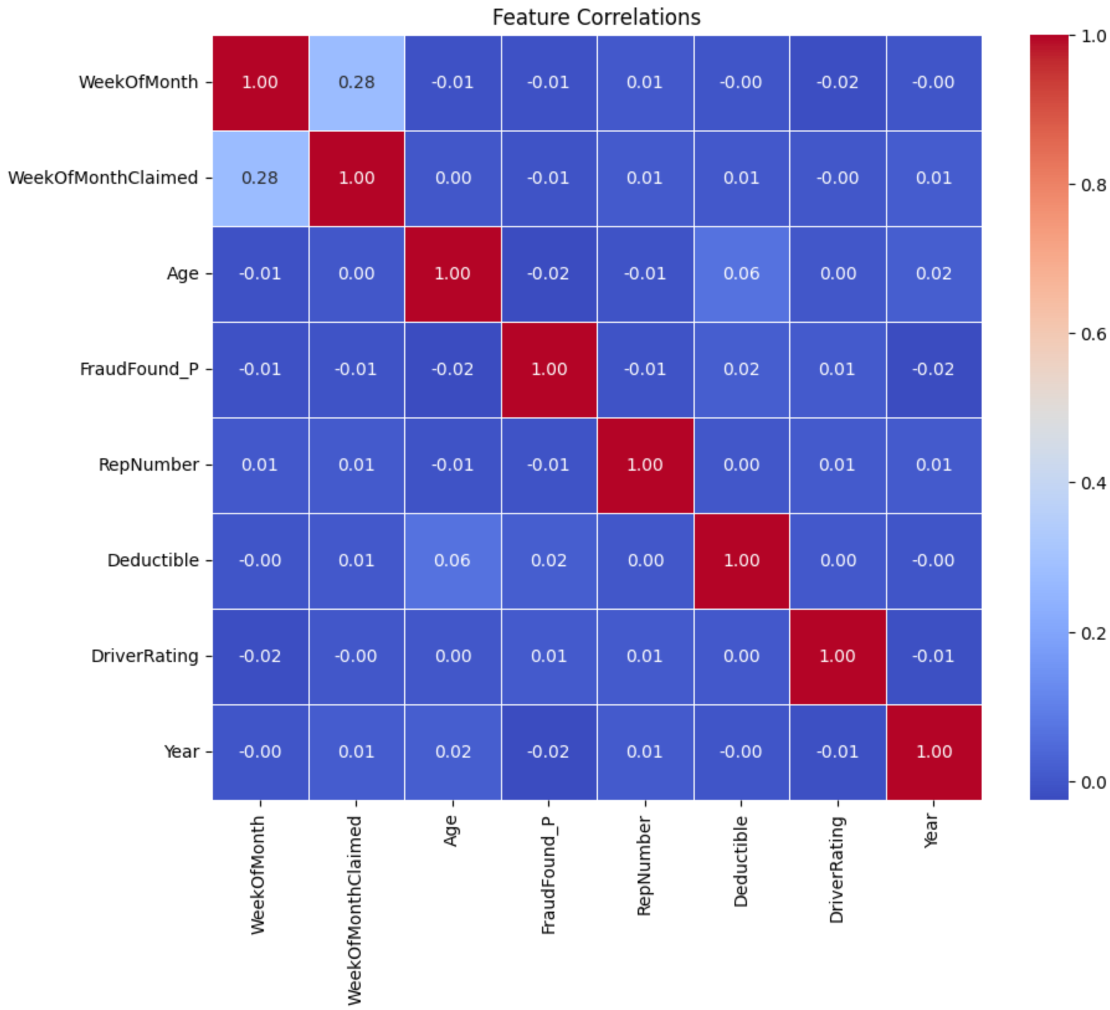
[96]:
```
# 4. Vehicle Price vs Fraud Rate
plt.figure(figsize=(10, 6))
fraud_rate = df.groupby('VehiclePrice')['FraudFound_P'].mean()
sns.barplot(x=fraud_rate.index, y=fraud_rate.values * 100,␣
    ↪palette='viridis',hue=fraud_rate.index, legend=False)
plt.title(f'{'VehiclePrice'} vs. Fraud Rate', fontsize=14)
plt.ylabel('Fraud Rate (%)', fontsize=12)
plt.xlabel('VehiclePrice', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.savefig(f'EDA_Charts/{'VehiclePrice'}_fraud_rate.png')
plt.show()
```



- Age distribution
- Policy type breakdown
- Geographic patterns

4. Correlation Analysis

[97]:
```
# Select numerical columns only
numerical_cols = df.select_dtypes(include=['int64', 'float64'])
numerical_cols = numerical_cols.drop('PolicyNumber',axis = 1)
corr_matrix = numerical_cols.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Feature Correlations')
plt.savefig('EDA_Charts/correlation_heatmap.png')
plt.show()
```



2 Correlation Analysis Report

[98]:
```
# Correlation Analysis Report
data = {'Column 1': ['WeekOfMonth', 'FraudFound_P', 'FraudFound_P',␣
↪'FraudFound_P', 'All Other Pairs'],
        'Column 2': ['WeekOfMonthClaimed', 'Age', 'Deductible', 'DriverRating',␣
↪'(Various)'],
        'Coefficient (r)': ['0.28', '0.022', '0.017', '-0.025', ' ± 0.01 or␣
↪less'],
        'Implication': ['Moderate temporal link', 'Very weak predictor', 'Very␣
↪weak predictor', 'Very weak predictor', 'Not significant']}
df_corr_summary = pd.DataFrame(data)
print("Concise Pairwise Correlation Analysis\n")
```

```
print(df_corr_summary.to_markdown(index=False, numalign="left",␣
↪stralign="left"))
```

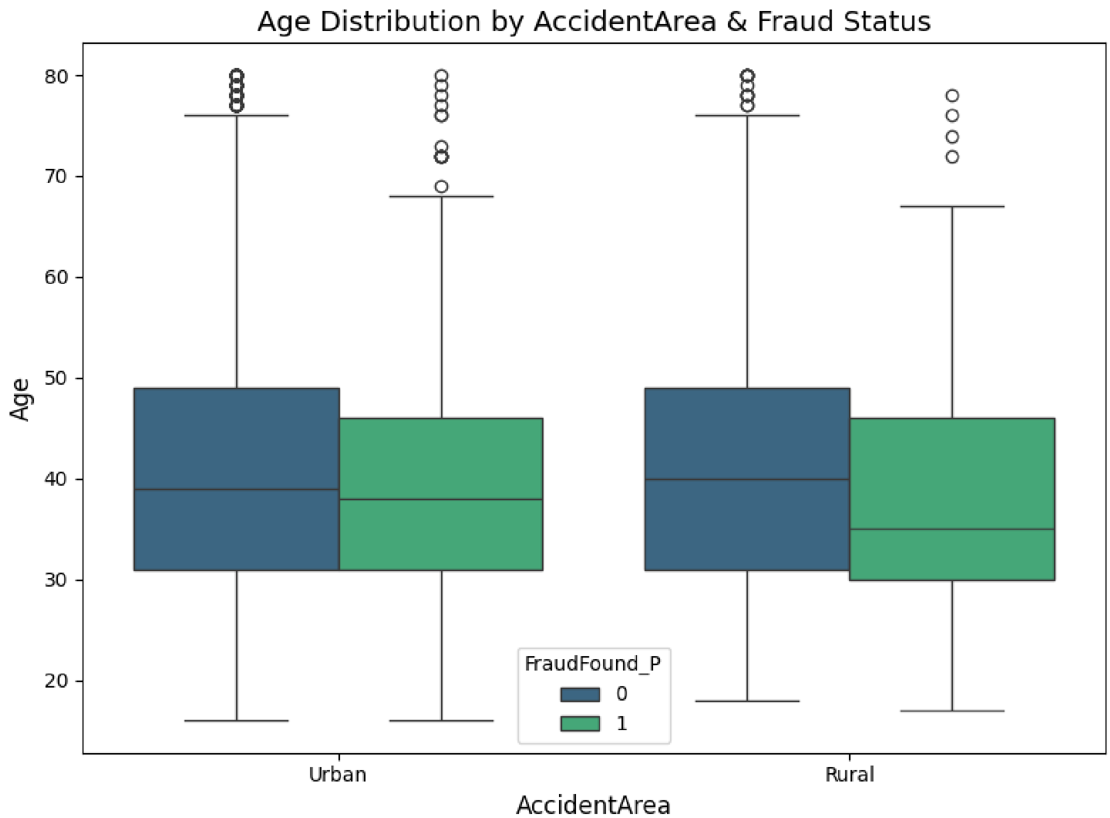
```
Concise Pairwise Correlation Analysis

| Column 1         | Column 2           | Coefficient (r)    | Implication
|
|:-----------------|:-------------------|:-------------------|:--------------------
----|
| WeekOfMonth      | WeekOfMonthClaimed | 0.28               | Moderate temporal
link |
| FraudFound_P     | Age                | 0.022              | Very weak predictor
|
| FraudFound_P     | Deductible         | 0.017              | Very weak predictor
|
| FraudFound_P     | DriverRating       | -0.025             | Very weak predictor
|
| All Other Pairs  | (Various)          |   ± 0.01 or less   | Not significant
|
```

5. Feature Relationships

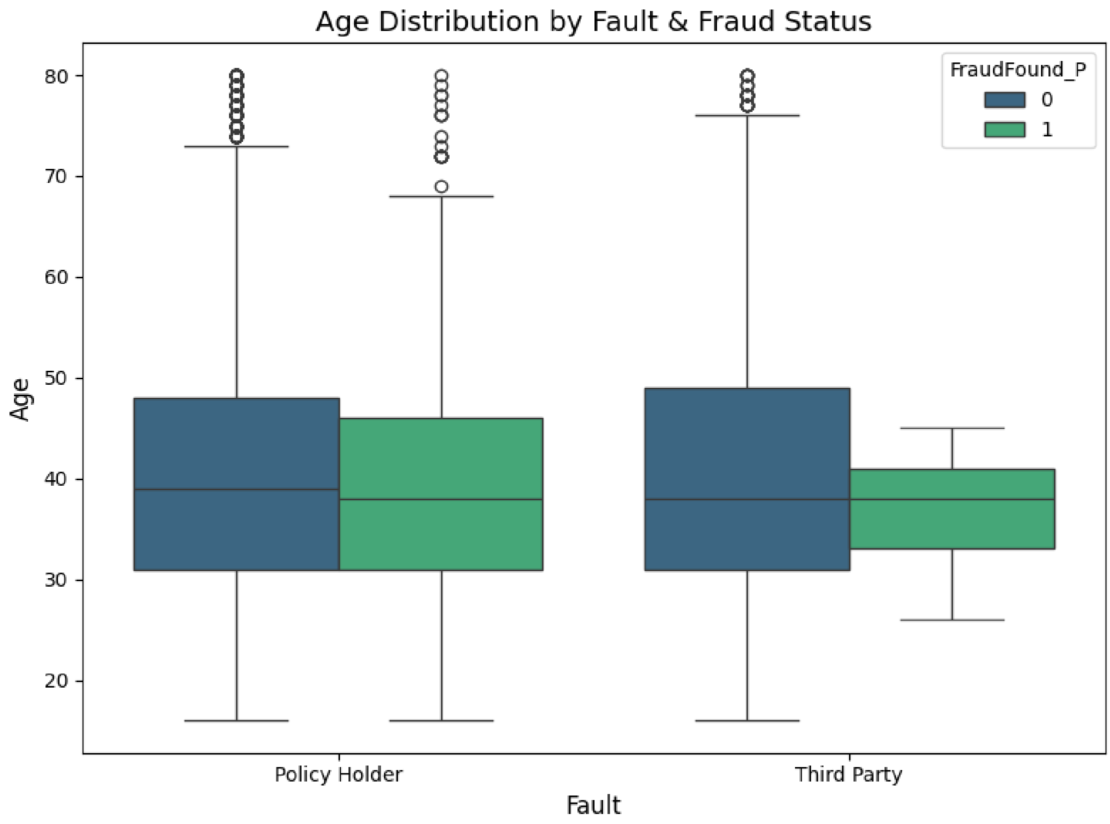
- Scatter plots for claim_amount vs age
- Box plots for categorical features vs fraud

2.0.1 Box plots for categorical features vs fraud

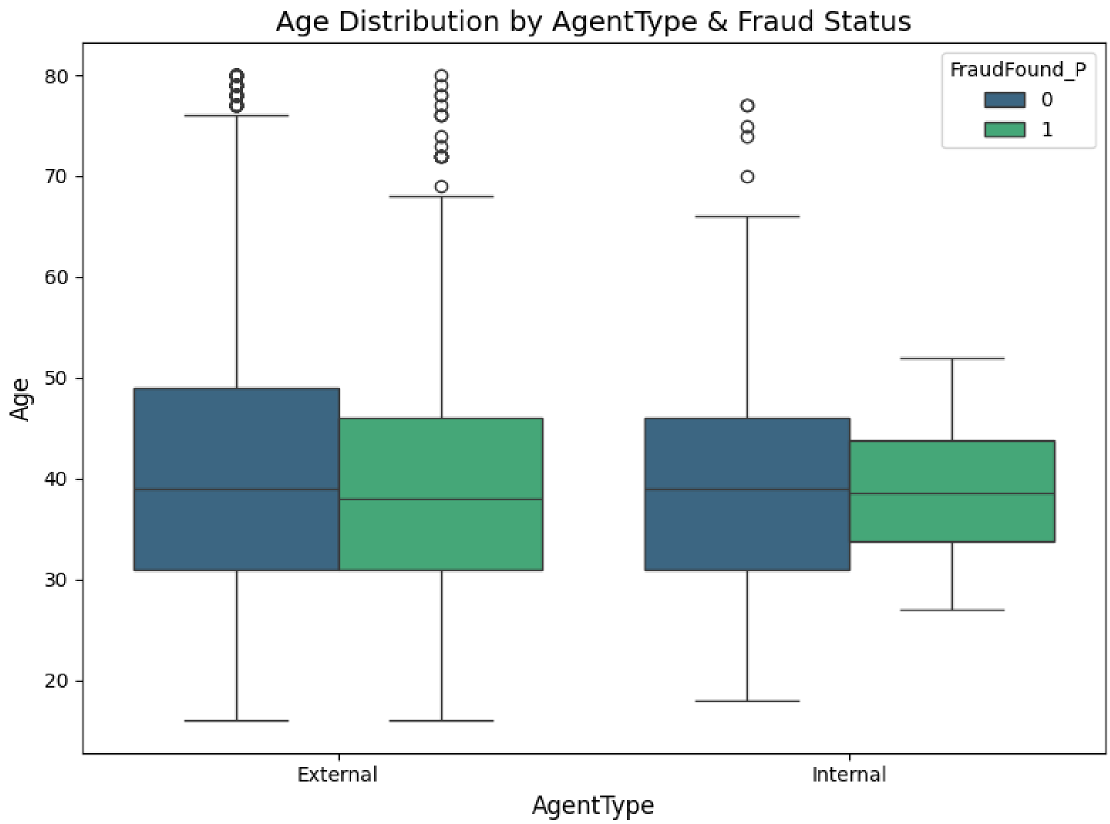
[109]:
```
# Age Distribution by AccidentArea & Fraud Status

plt.figure(figsize=(8, 6))
colname='AccidentArea'
sns.boxplot(x=colname, y='Age', hue='FraudFound_P', data=df, palette='viridis')
plt.title(f'Age Distribution by {colname} & Fraud Status', fontsize=14)
plt.ylabel('Age', fontsize=12)
plt.xlabel(colname, fontsize=12)
plt.tight_layout()
plt.savefig('EDA_Charts/boxplot_accident_area.png')
plt.show()
```



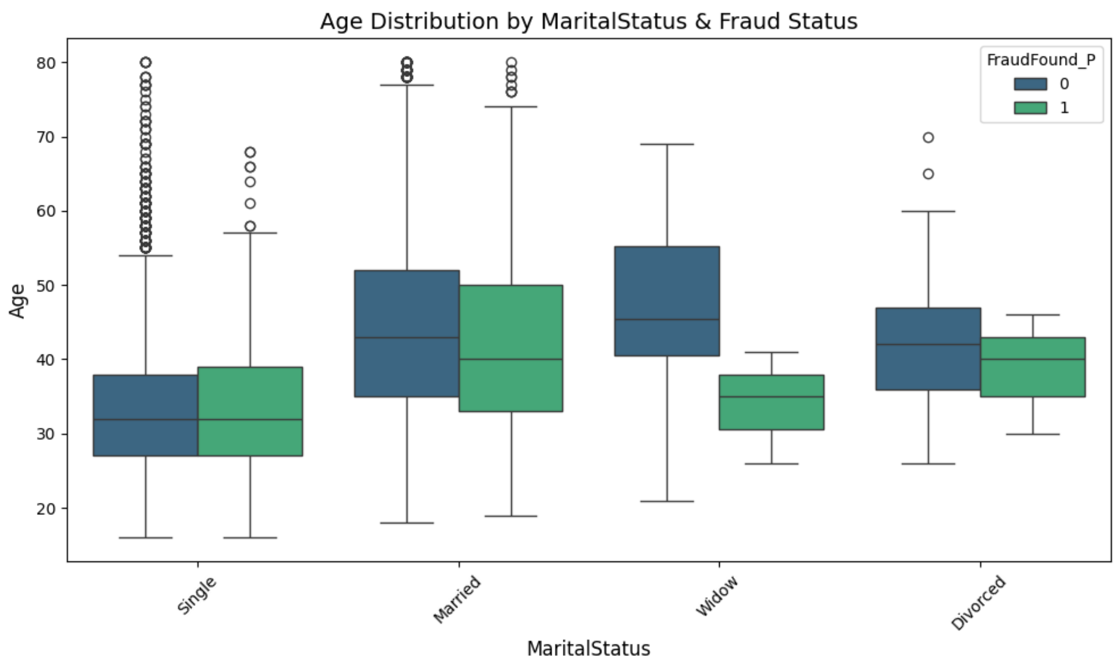
[103]:
```
colname='Fault'
plt.figure(figsize=(8, 6))
sns.boxplot(x=colname, y='Age', hue='FraudFound_P', data=df, palette='viridis')
plt.title(f'Age Distribution by {colname} & Fraud Status', fontsize=14)
plt.ylabel('Age', fontsize=12)
plt.xlabel(colname, fontsize=12)
plt.tight_layout()
plt.savefig('EDA_Charts/boxplot_fault.png')
plt.show()
```



[104]:
```
colname='AgentType'
plt.figure(figsize=(8, 6))
sns.boxplot(x=colname, y='Age', hue='FraudFound_P', data=df, palette='viridis')
plt.title(f'Age Distribution by {colname} & Fraud Status', fontsize=14)
plt.ylabel('Age', fontsize=12)
plt.xlabel(colname, fontsize=12)
plt.tight_layout()
plt.savefig('EDA_Charts/boxplot_agent_type.png')
plt.show()
```

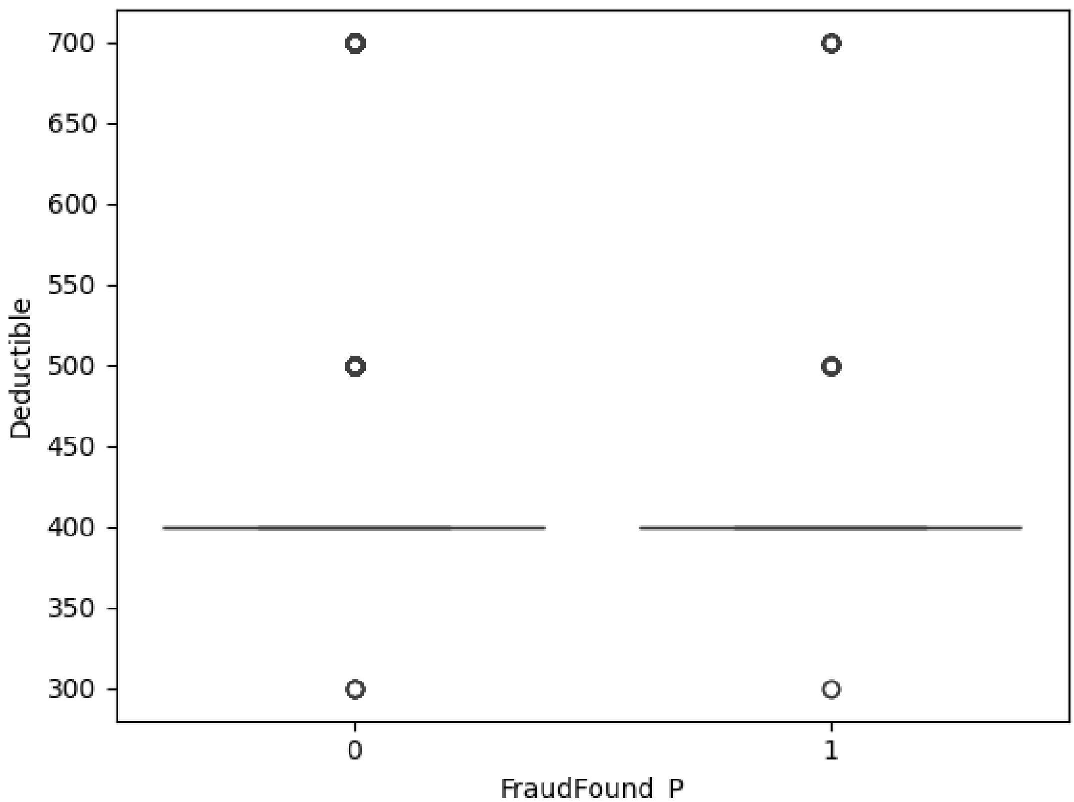


[105]:
```
colname='MaritalStatus'
plt.figure(figsize=(10, 6))
sns.boxplot(x=colname, y='Age', hue='FraudFound_P', data=df, palette='viridis')
plt.title(f'Age Distribution by {colname} & Fraud Status', fontsize=14)
plt.ylabel('Age', fontsize=12)
plt.xlabel(colname, fontsize=12)
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('EDA_Charts/boxplot_marital_status.png')
plt.show()
```



[106]:
```
sns.boxplot(x='FraudFound_P', y='Deductible', data=df)
```

[106]: <Axes: xlabel='FraudFound_P', ylabel='Deductible'>



6. Create EDA Report

- Document findings in notebook
- Note: Which features seem important?
- Note: Any data quality issues?
- Note: Class imbalance ratio

[107]:
```python
# Select the categorical columns identified as important
categorical_features = ['Fault', 'AgentType', 'AccidentArea', 'MaritalStatus']
```

3 Exploratory Data Analysis (EDA) Report

3.1 1. Data Structure & Quality

Dataset Size: 15,419 rows, 33 columns.

Completeness: 0 null values found across the dataset. No imputation required.

Data Types: - 9 Numeric columns (e.g., Age, Deductible, DriverRating).

24 Categorical columns (e.g., Make, AccidentArea, Fault).

Recommendation: Drop 'PolicyNumber' to prevent multicollinearity in modeling.

3.2 2. Class Imbalance Analysis

Target Variable: FraudFound_P (0 = Genuine, 1 = Fraud)

Distribution:

`Genuine Claims (0): ~94% (14,496 records)`

`Fraudulent Claims (1): ~6% (923 records)`

Imbalance Ratio: Approximately 15.7 : 1

Impact:Accuracy is a misleading metric; models will require resampling (SMOTE) or specific evaluation metrics (F1-Score, ROC-AUC).

3.3 3. Feature Importance Findings

3.3.1 Strong Predictors (Key Categorical Features)

Categorical features showed the strongest separation between fraud and genuine cases:

Fault: The single strongest indicator. Claims where the Policy Holder is at fault show a significantly higher fraud rate (>10%) compared to Third Party fault (~1%).

Vehicle Price: Positive trend observed; vehicles in higher price brackets (>$69k) correlate with higher fraud rates.

Age of Policy Holder: Younger policyholders (26–30 range) exhibit higher fraud frequency compared to older groups.

Agent Type: Claims filed through 'Internal' agents have a noticeably higher fraud rate than 'External' agents.

Accident Area: 'Urban' areas show a higher incidence of fraud compared to 'Rural' areas.

3.3.2 Weak Predictors (Numeric Features)

Numeric features generally showed poor linear correlation with fraud:

Deductible: Distribution is nearly identical for both Fraud and Genuine cases (Median $400).

Driver Rating: Correlation with fraud is negligible (-0.02).

Age (Numeric): Weak linear correlation (0.02). However, when treated as categorical bins (Age Groups), it becomes predictive.

3.4 4. Outlier Analysis

Age: Outliers detected via IQR analysis in specific sub-groups (e.g., 'Widow' and 'Single' marital statuses), though values remain within realistic human age ranges.

Deductible: Minor outliers observed at $500 and $700 levels, but these do not distinctly separate fraud classes.