Assignment 2 - Correlation

INSTRUCTIONAL DETAILS

In Python, there are a few ways to calculate the correlation between two variables. One way is to use the pearsonr function from the scipy.stats module. This function calculates the Pearson correlation coefficient, which is a measure of the linear relationship between two variables.

Here's an example of how you could use the pearsonr function to calculate the correlation between two variables in Python:

```
from scipy.stats import pearsonr

x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

corr, p_value = pearsonr(x, y)

print(f'Correlation: {corr}')
```

This would output Correlation: 1.0, indicating that there is a strong positive correlation between x and y.

Another way to calculate the correlation between two variables is to use the corr method of a pandas dataframe. This method calculates the Pearson correlation coefficient by default, but it can also calculate other types of correlations, such as the Spearman rank correlation coefficient or the Kendall rank correlation coefficient.

Here's an example of how you could use the corr method to calculate the Pearson correlation coefficient between two variables in a pandas dataframe:

```
import pandas as pd

df = pd.DataFrame({'x': [1, 2, 3, 4, 5], 'y': [2, 4, 6, 8, 10]})

corr = df.corr()

print(corr)
```

This would output a pandas dataframe with the correlation coefficients between all pairs of variables in the dataframe, like this:

```
     x         y
x  1.000000  1.000000
y  1.000000  1.000000
```

You can also use the numpy module to calculate the correlation between two variables. The numpy.corrcoef function calculates the Pearson correlation coefficient by default, but it can also calculate other types of correlations if you specify a different correlation coefficient.

Here's an example of how you could use the numpy.corrcoef function to calculate the Pearson correlation coefficient between two variables:

```
import numpy as np

x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

corr = np.corrcoef(x, y)

print(corr)
```

This would output a 2x2 matrix with the correlation coefficients between x and y, like this:

```
[[1. 1.]
 [1. 1.]]
```

Here's an example of a correlation matrix using Python:

```
import pandas as pd

# Sample data
data = {'A': [1, 2, 3, 4, 5],
        'B': [2, 4, 6, 8, 10],
        'C': [3, 6, 9, 12, 15]}
df = pd.DataFrame(data)

# Calculate the correlation matrix
corr_matrix = df.corr()

# Print the correlation matrix
print(corr_matrix)
```

Here's an example of a heat map (for the correlation matrix) using Python:

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

# Sample data
data = {'A': [1, 2, 3, 4, 5],
        'B': [2, 4, 6, 8, 10],
        'C': [3, 6, 9, 12, 15]}
df = pd.DataFrame(data)

# Calculate the correlation matrix
corr_matrix = df.corr()

# Create the heatmap
sns.heatmap(corr_matrix, annot=True)

# Show the plot
plt.show()
```

Here's an example of a well labeled XY scatter using Python:

```
import matplotlib.pyplot as plt
import numpy as np

# Create some example data
x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

# Create the scatter plot
plt.scatter(x, y)

# Add a trendline
z = np.polyfit(x, y, 1)
p = np.poly1d(z)
plt.plot(x,p(x),"r--")

# Add title and axis labels
plt.title('Scatter Plot with Trendline')
plt.xlabel('X')
plt.ylabel('Y')

# Show the plot
plt.show()
```

BACKGROUND

Correlation is a statistical measure that describes the relationship between two variables. It is a useful tool for businesses because it allows them to understand how two variables are related to each other.

For example, a business might use correlation to understand the relationship between advertising spend and sales. If there is a strong positive correlation between these two variables, it might suggest that increasing advertising spend leads to increased sales. On the other hand, if there is a strong negative correlation, it might suggest that increasing advertising spend leads to decreased sales.

Correlation can also be used to identify patterns and trends in data that might not be immediately apparent. For example, a business might use correlation to identify a relationship between customer satisfaction and repeat business. If there is a strong positive correlation between these two variables, it might suggest that improving customer satisfaction leads to increased repeat business.

Overall, correlation is an important tool for businesses because it helps them understand the relationships between different variables and how they might be related to important outcomes. This can inform decision-making and help businesses achieve their goals.

RESEARCH QUESTION

Is there a relationship (significant, or otherwise) between the number of claims, number of properties, and total amount paid for flood damage by UK postal codes?

Data analysis can be a valuable tool in the flood insurance industry, as it can help insurers better understand and assess flood risk, and can be used to develop more accurate flood risk models. Here are a few ways in which data analysis is used in the flood insurance industry:

Flood risk assessment: Insurers can use data analysis to assess the flood risk of a particular property or area. This can involve analyzing data on factors such as the elevation of the property, the proximity to bodies of water, and the likelihood of heavy rain or snowmelt.

Pricing: Insurers can use data analysis to help determine the premiums for flood insurance policies. This may involve analyzing data on past flood events, the likelihood of future flood events, and the potential cost of damages.

Claim processing: Data analysis can be used to help streamline the process of processing flood insurance claims. This may involve analyzing data on the damages sustained, the cost of repairs, and the policy coverage.

Risk management: Insurers can use data analysis to identify trends and patterns in flood risk and to develop strategies for managing and mitigating that risk. This may involve analyzing data on flood prevention and mitigation measures, such as levees and flood control structures.

REQUIREMENTS FOR SUBMISSION

See “Assignment 1 - Descriptives” for submission requirements.

FORMATTING

See “Assignment 1 - Descriptives” for a detailed list of assignment formatting guidelines. Also, assignment formatting guidelines can be found in the course document cache.

DATASET DETAILS (all data sets can be found here or here)

flood.csv

DATASET FIELDS

uk_post_codes
claims
properties
amount_paid