HW_02 by Andrew Lee

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?

Load the data from google drive.

In [2]:

```
from google.colab import drive
drive.mount("/content/gdrive")
```

```
Mounted at /content/gdrive
```

In [3]:

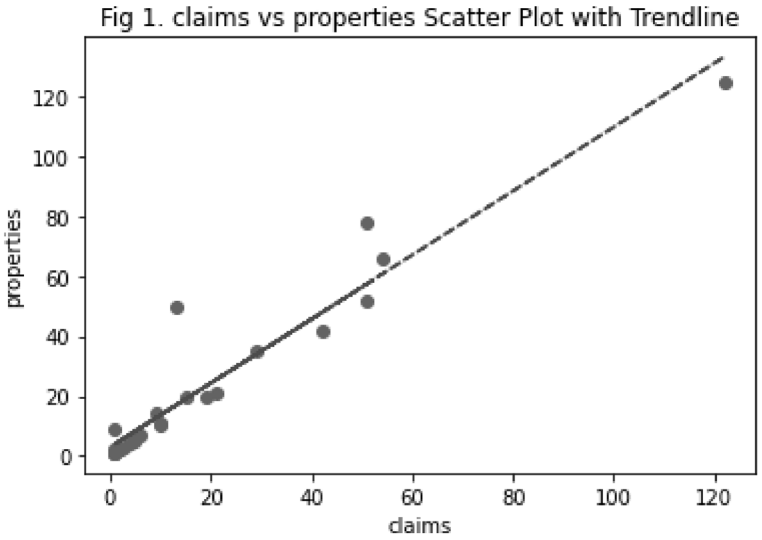
```
import pandas as pd

# import data from csv

df = pd.read_csv('/content/gdrive/MyDrive/Colab Notebooks/dataset/wentworth_applied_ana
lytics - flood.csv.csv')

print(df.head())
print(df.columns)
```

```
  uk_post_codes  claims  properties  amount_paid
0         HD6 1      19          20        92180
1         HD6 2       1           1         5000
2         HD6 3       1           1         5000
3         HD6 4       5           5        12758
4         HX2 6      15          20        95384
Index(['uk_post_codes', 'claims', 'properties', 'amount_paid'], dtype='obj
ect')
```

Corrlation between the number claims and the number properties

In [9]:

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import pearsonr

x = df["claims"]
y = df["properties"]
# 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('Fig 1. claims vs properties Scatter Plot with Trendline')
plt.xlabel('claims')
plt.ylabel('properties')

# Show the plot
plt.show()

r, p = pearsonr(x, y)
print("Pearson correlation coefficient: ", r)
print("p-value: ", p)
alpha = 0.05
if p > alpha:
    print("The correlation is not statistically significant.")
else:
    print("The correlation is statistically significant.")
```



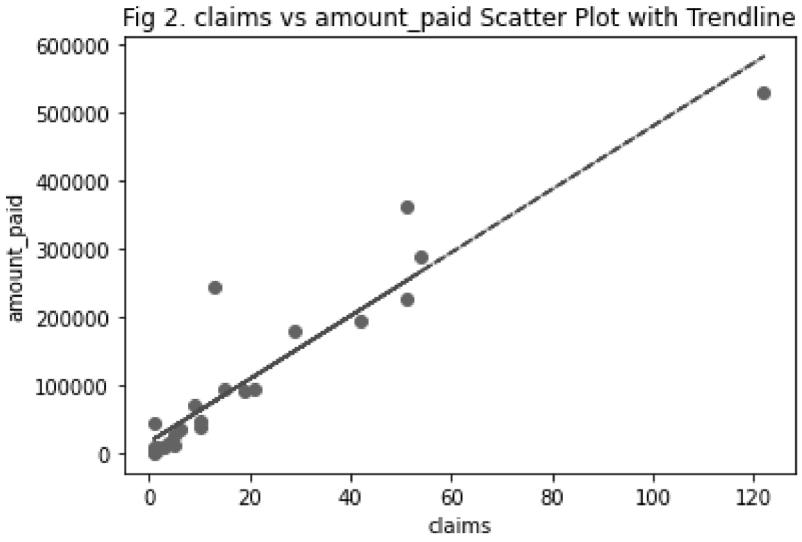

```
Pearson correlation coefficient:  0.9583455940876426
p-value:  1.4551939378821306e-14
The correlation is statistically significant.
```

From Fig 1. Scatter Plot, We can see there is a positive trend between the number of claims and properties. As calims increase, properties increase. The Pearson correlation coefficient is around 0.958, close to 1. The p-value is 1.455e^(-14) which is lower 0.05 and we can say the correlation is statistically significant.

Corrlation between the number claims and the total amount of paid

In [10]:

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import pearsonr

x = df["claims"]
y = df["amount_paid"]
# 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('Fig 2. claims vs amount_paid Scatter Plot with Trendline')
plt.xlabel('claims')
plt.ylabel('amount_paid')

# Show the plot
plt.show()

r, p = pearsonr(x, y)
print("Pearson correlation coefficient: ", r)
print("p-value: ", p)
alpha = 0.05
if p > alpha:
    print("The correlation is not statistically significant.")
else:
    print("The correlation is statistically significant.")
```



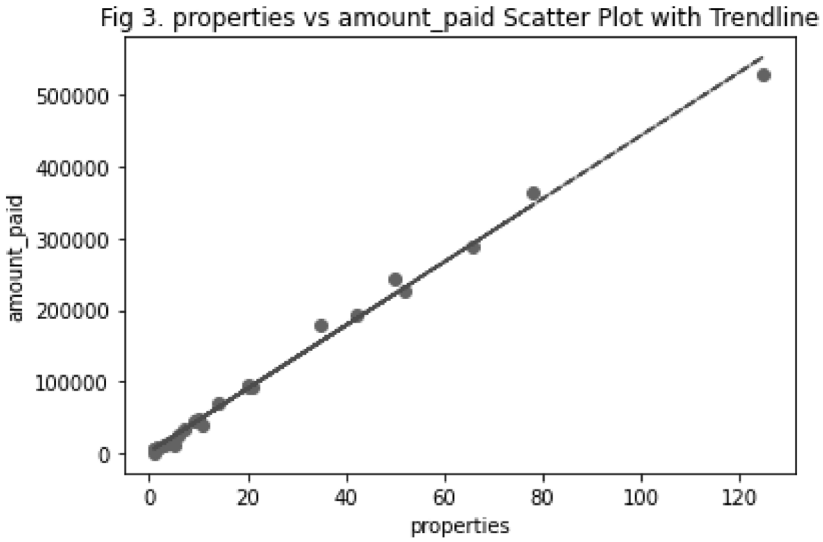

```
Pearson correlation coefficient:  0.9410232215154994
p-value:  8.632121195909054e-13
The correlation is statistically significant.
```

From Fig 2. Scatter Plot, We can see there is a positive trend between the number of claims and the total amount of paid. As calims increase, total amount of paid increase. The Pearson correlation coefficient is around 0.941, close to 1. The p-value is 8.632e^(-13) which is lower 0.05 and we can say the correlation is statistically significant.

Corrlation between the number properties and the total amount of paid

In [11]:

```
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import pearsonr

x = df["properties"]
y = df["amount_paid"]
# 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('Fig 3. properties vs amount_paid Scatter Plot with Trendline')
plt.xlabel('properties')
plt.ylabel('amount_paid')

# Show the plot
plt.show()

r, p = pearsonr(x, y)
print("Pearson correlation coefficient: ", r)
print("p-value: ", p)
alpha = 0.05
if p > alpha:
    print("The correlation is not statistically significant.")
else:
    print("The correlation is statistically significant.")
```




```
Pearson correlation coefficient:  0.997251064608996
p-value:  1.2122201395931224e-28
The correlation is statistically significant.
```

From Fig 3. Scatter Plot, We can see there is a positive trend between the number of properties and the total amount of paid. As properties increase, total amount of paid increase. The Pearson correlation coefficient is around 0.941, close to 1. The p-value is 1.212e^(-28) which is lower 0.997 and we can say the correlation is statistically significant.

Base on all the scatter plot infomration, I would say the number of properties and the total amount of paid has the most significant and positive relationship. There are 3 reason:

- it has the least p-value.
- the corrlation coefficient is cloest to 1.
- there is no outliers in scatter plot and most of the points are on the line.

However,since the dataset size is small, I could say the statements are only true for current data. Using above statment, we can try to apply linear regression to predition.