EXPERIMENT NO- 4

AIM: To study Exploratory Data Analysis and visualization of Social Media Data for business.

RESOURCES REQUIRED: Windows/MAC/Linux O.S, Compatible version of Python.

THEORY:

What is Exploratory Data Analysis?

We can define exploratory data analysis as the essential data investigation process before the formal analysis to spot patterns and anomalies, discover trends, and test hypotheses with summary statistics and visualizations. It gives an idea about the data we will be digging deep into while analyzing. It aids in formulating how we can handle data during analysis, like choosing models, handling outliers, deciding model accuracy parameters, etc. Visualization helps to infer insights easily from massive datasets.

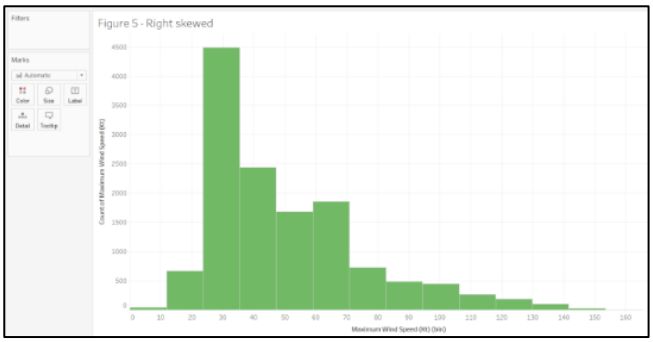
Types of Exploratory Data Analysis

1. Univariate Plots

Univariate plots show the frequency or the distribution shape of a variable.

2. Histograms

Histograms are two-dimensional plots in which the x-axis divide into a range of numerical bins or time intervals. The y-axis shows the frequency values, which are counts of occurrences of values for each bin. Bar graphs have gaps between the bars to indicate that they compare distinct groups, but there are no gaps in histograms. Hence, They tell us if the distribution is left/positively skew (most of the data falls to the right side), right/negatively skewed (most of the data falls to the left side), bi-modal (graphs having two distinct peaks), normal (perfectly symmetrical without skew), or uniform (almost all the bins have similar frequency).



Probability Distribution Plots

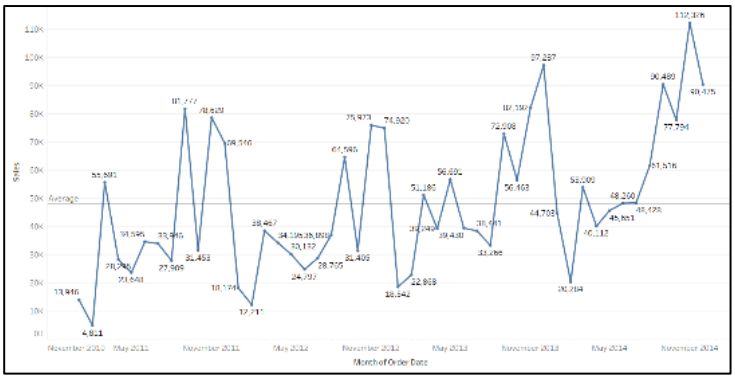
Probability distributions are mathematical functions that describe all the possible values that a random variable can assume within a given range. They help model random phenomena, allowing us in order to estimate the probability of a particular event. This type of distribution is helpful to know the likely outcomes and the spread of potential values.

For a single random variable, probability distributions can be divided into two types:

1. Discrete Probability Distributions for Discrete Variables
2. Binomial Distribution

Run Sequence Plots

A run chart, also known as a run-sequence plot, displays observed data in a time sequence. So, Often, the data displayed represents some aspect of a business process's output or performance. It is, therefore, a form of a line chart. They are often analyzed in order to locate anomalies in data that suggest shifts in a process over time. Changes in location and scale and outliers can easily be detected.

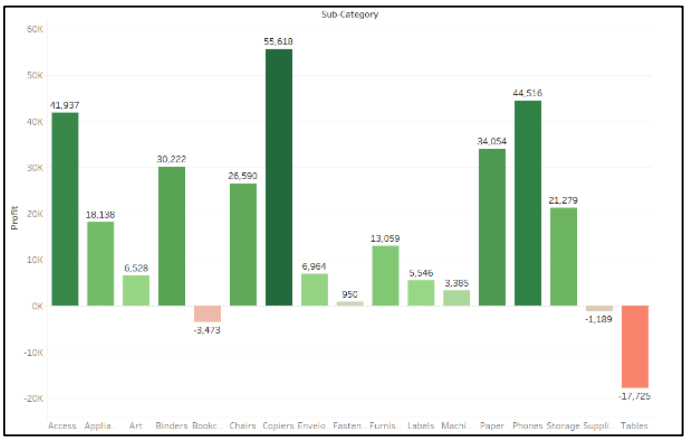


Bivariate Plots

Bivariate plots display the relationship between two variables in exploratory data analysis.

Bar Graphs

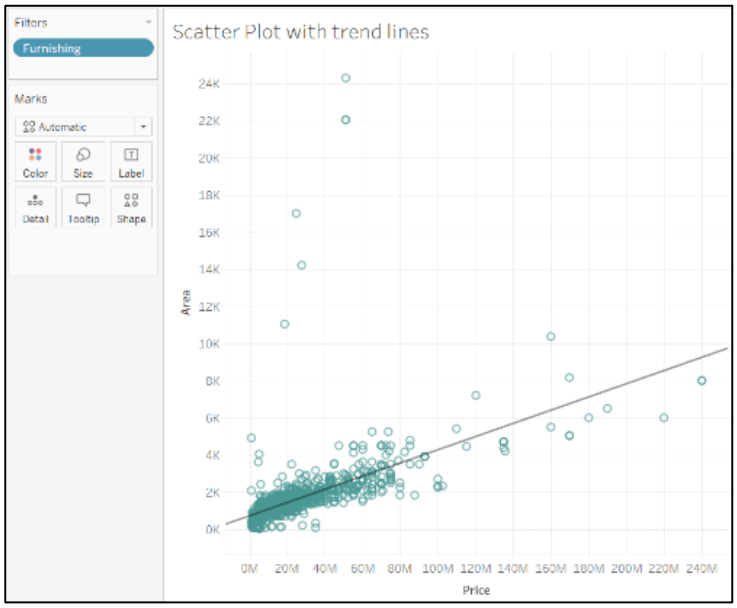
Bar charts can be used to compare nominal or ordinal data. They are helpful for recognizing trends.



Scatter Plots

Scatter plots are commonly used in statistical analysis in order to visualize numerical relationships. So, They are use in order to determine whether two measures are correlate by plotting them on the x and y-axis. They are suitable for recognizing trends.

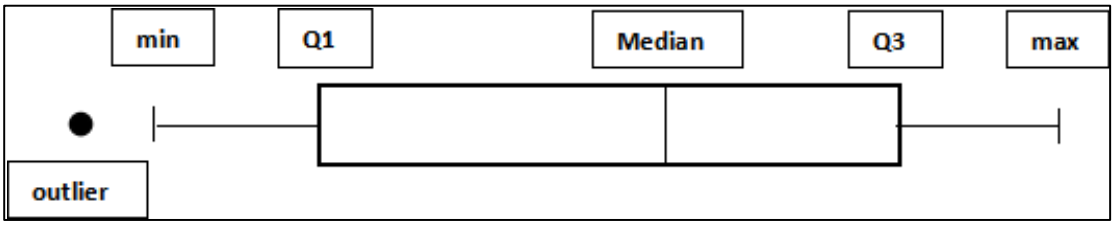
For instance, you can see a scatter plot of two measures in the figure – the house's area against price and the trend line. The data points are concentrated in the lower price and lower area range. A few outliers are indicating larger area houses available for lower prices.



Box Plots

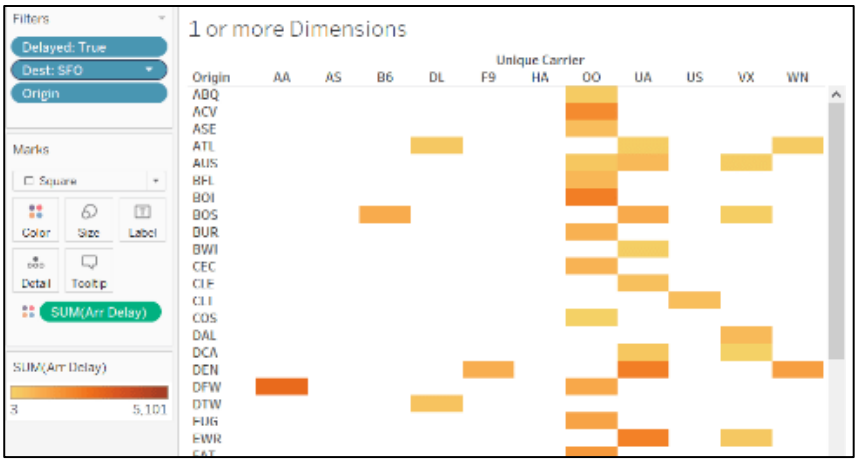
These charts show the distribution of values along an axis. Rectangular boxes are used in order to bucket the data, giving us an idea of how the data points are spread out. These boxes are also called quartiles which represent a quarter of a data set. Boxes can be drawn vertically or horizontally.

Box plots are suitable for identifying outliers. The below figure shows the structure of a box plot.



Heat Maps

For instance, correlation heat maps show the interrelationship between variables—areas as shaded as per the data's values. So, Color differences can easily spot similar and different values and make sense of the data variation. They are usually helpful when you have a large amount of data. They are used during A/B testing to see which parts of a web page are accessed by users on a website.



CONCLUSION: Hence, we have successfully studied Exploratory Data Analysis and visualization of Social Media Data for business.

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
post_df = pd.read_csv('post.csv',index_col=['id'])
comment_df = pd.read_csv('comment.csv',index_col=['comment_id'])
```

In [3]:
```
post_df.head()
```

Out[3]:

id	author	author_id	total_comments	upvote	post_type	ti
12glkw4	TheBodyPolitic1	von3w6y2	319	591	top week	Why did Python becor popular ur long a
12hj9oc	MetonymyQT	shnqm	80	527	top week	Free cours Build a mode API w FastAPI a
12egsoz	2broke2code	rs4dqilj	105	467	top week	I trainec RoastBot >120,000 fac and >0
12ffsif	midnitte	3gad9	64	378	top week	EP 684: A P Interpreter G Accept
12fzdu2	aeluro1	88efmodhn	12	370	top week	Comprehensi Reddit Sav Pos Downloade

In [4]:
```
comment_df.head()
```

Out[4]:

comment_id	post_id	parent_id	body	created_on	upvotes	author_id	author_
jfkvwx0	12glkw4	t3_12glkw4	Hardware wasn't ready for Python in that time	2023-04-09 15:27:18	5	22vat21u	Du
jfnha98	12glkw4	t3_12glkw4	Because Python was developed with the conceit ...	2023-04-10 02:55:16	2	4wtjvsh6	FredVIII
jflnf2e	12glkw4	t3_12glkw4	Perl was *the* scripting language in the early...	2023-04-09 18:37:04	2	4i9hp	tom
jflbch7	12glkw4	t3_12glkw4	I was a web developer between 2000 and 2010, a...	2023-04-09 17:13:45	2	8hi6986p	As Dress
jfmk972	12glkw4	t1_jflbch7	Adding on to this from my POV, early web in th...	2023-04-09 22:33:38	3	380he	snap

Analysing 1st post comments

In [5]:
```
fp_comments = comment_df.loc[comment_df['post_id']=='12glkw4']
fp_comments.head()
```

Out[5]:

comment_id	post_id	parent_id	body	created_on	upvotes	author_id	author_
jfkvwx0	12glkw4	t3_12glkw4	Hardware wasn't ready for Python in that time	2023-04-09 15:27:18	5	22vat21u	Du
jfnha98	12glkw4	t3_12glkw4	Because Python was developed with the conceit ...	2023-04-10 02:55:16	2	4wtjvsh6	FredVIII
jflnf2e	12glkw4	t3_12glkw4	Perl was *the* scripting language in the early...	2023-04-09 18:37:04	2	4i9hp	tom
jflbch7	12glkw4	t3_12glkw4	I was a web developer between 2000 and 2010, a...	2023-04-09 17:13:45	2	8hi6986p	As Dress
jfmk972	12glkw4	t1_jflbch7	Adding on to this from my POV, early web in th...	2023-04-09 22:33:38	3	380he	snap

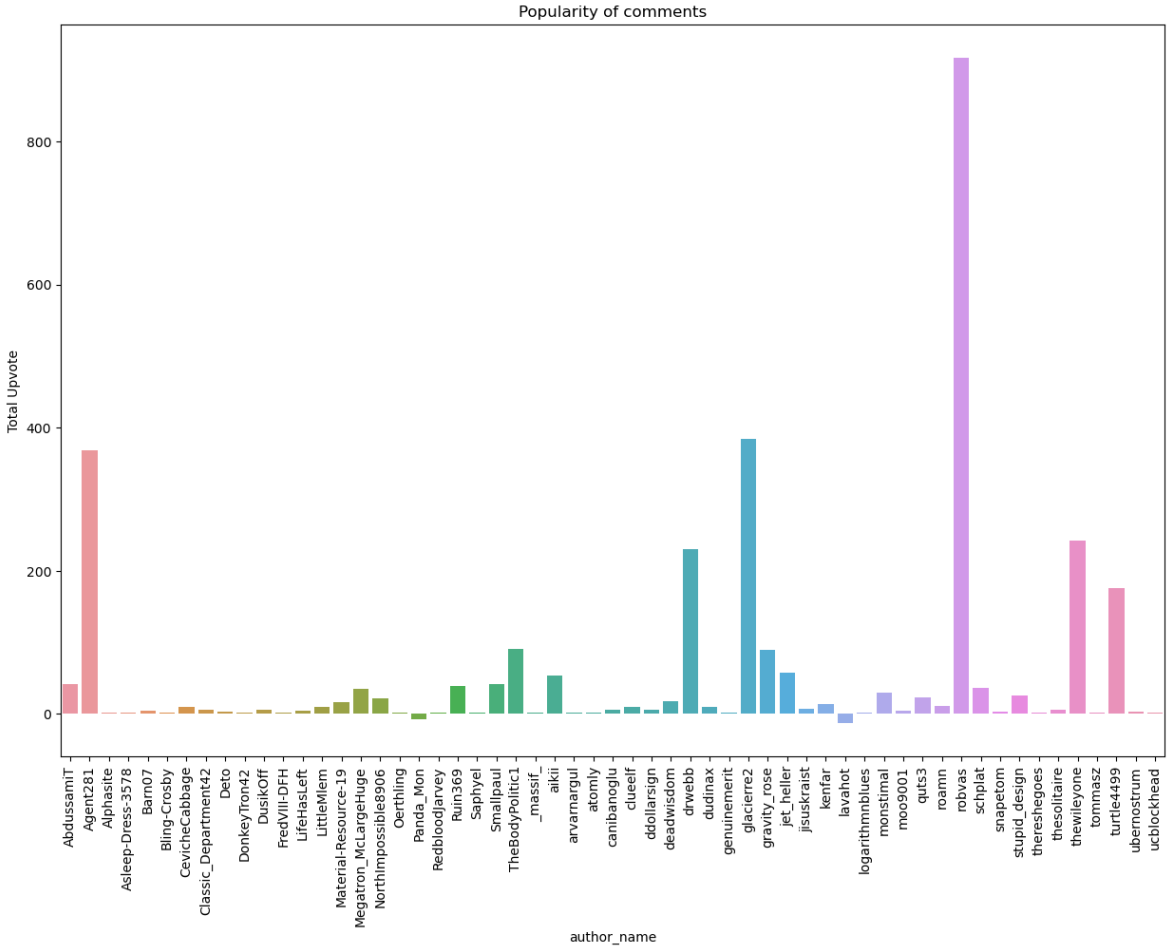
In [6]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [7]:
```
upvote_rec = fp_comments.groupby(fp_comments['author_name']).apply(lambda
    "total_upvote":x['upvotes'].sum(),
    'total_comments':len(x)
})
upvote_rec
```

Out[7]:
```
author_name
AbdussamiT                {'total_upvote': 42, 'total_comments': 2}
Agent281                 {'total_upvote': 368, 'total_comments': 1}
Alphasite                  {'total_upvote': 2, 'total_comments': 1}
Asleep-Dress-3578          {'total_upvote': 2, 'total_comments': 1}
Barn07                     {'total_upvote': 4, 'total_comments': 1}
Bling-Crosby               {'total_upvote': 2, 'total_comments': 1}
CevicheCabbage             {'total_upvote': 9, 'total_comments': 1}
Classic_Department42       {'total_upvote': 6, 'total_comments': 1}
Deto                       {'total_upvote': 3, 'total_comments': 1}
DonkeyTron42               {'total_upvote': 2, 'total_comments': 1}
DusikOff                   {'total_upvote': 5, 'total_comments': 1}
FredVIII-DFH               {'total_upvote': 2, 'total_comments': 1}
LifeHasLeft                {'total_upvote': 4, 'total_comments': 2}
LittleMlem                 {'total_upvote': 9, 'total_comments': 1}
Material-Resource-19      {'total_upvote': 16, 'total_comments': 1}
Megatron_McLargeHuge      {'total_upvote': 35, 'total_comments': 1}
NorthImpossible8906       {'total_upvote': 21, 'total_comments': 1}
Oerthling                  {'total_upvote': 1, 'total_comments': 1}
Panda_Mon                 {'total_upvote': -8, 'total_comments': 1}
RedbloodJarvey             {'total_upvote': 1, 'total_comments': 1}
Ruin369                   {'total_upvote': 39, 'total_comments': 1}
Saphyel                    {'total_upvote': 2, 'total_comments': 1}
Smallpaul                 {'total_upvote': 41, 'total_comments': 1}
TheBodyPolitic1           {'total_upvote': 90, 'total_comments': 3}
_massif_                   {'total_upvote': 1, 'total_comments': 1}
aikii                     {'total_upvote': 53, 'total_comments': 1}
arvarnargul                {'total_upvote': 1, 'total_comments': 1}
atomly                     {'total_upvote': 1, 'total_comments': 1}
canibanoglu                {'total_upvote': 5, 'total_comments': 1}
clueelf                   {'total_upvote': 10, 'total_comments': 1}
ddollarsign                {'total_upvote': 5, 'total_comments': 1}
deadwisdom                {'total_upvote': 18, 'total_comments': 1}
drwebb                   {'total_upvote': 230, 'total_comments': 1}
dudinax                    {'total_upvote': 9, 'total_comments': 1}
genuinemerit               {'total_upvote': 2, 'total_comments': 1}
glacierre2               {'total_upvote': 384, 'total_comments': 1}
gravity_rose              {'total_upvote': 89, 'total_comments': 1}
jet_heller                {'total_upvote': 57, 'total_comments': 1}
jisuskraist                {'total_upvote': 7, 'total_comments': 1}
kenfar                    {'total_upvote': 13, 'total_comments': 1}
lavahot                  {'total_upvote': -13, 'total_comments': 1}
logarithmnblues            {'total_upvote': 1, 'total_comments': 1}
monstimal                 {'total_upvote': 29, 'total_comments': 1}
moo9001                    {'total_upvote': 4, 'total_comments': 1}
quts3                     {'total_upvote': 23, 'total_comments': 2}
roamn                     {'total_upvote': 11, 'total_comments': 1}
robvas                   {'total_upvote': 917, 'total_comments': 1}
schplat                   {'total_upvote': 36, 'total_comments': 1}
snapetom                   {'total_upvote': 3, 'total_comments': 1}
stupid_design             {'total_upvote': 26, 'total_comments': 1}
thereshegoes               {'total_upvote': 2, 'total_comments': 1}
thesolitaire               {'total_upvote': 6, 'total_comments': 1}
thewileyone              {'total_upvote': 242, 'total_comments': 1}
tommasz                    {'total_upvote': 2, 'total_comments': 1}
turtle4499               {'total_upvote': 175, 'total_comments': 1}
ubernostrum                {'total_upvote': 3, 'total_comments': 1}
ucblockhead                {'total_upvote': 1, 'total_comments': 1}
dtype: object
```

Most upvotes users in comments

In [8]:
```
fig,ax = plt.subplots(1,1,figsize=(15,10))
plt.xticks(rotation=90)
ax.set_xlabel('author_name')
ax.set_ylabel('Total Upvote')
ax.set_title('Popularity of comments')
sns.barplot(x=upvote_rec.index,y=[data['total_upvote'] for data in upvote
plt.show()
```



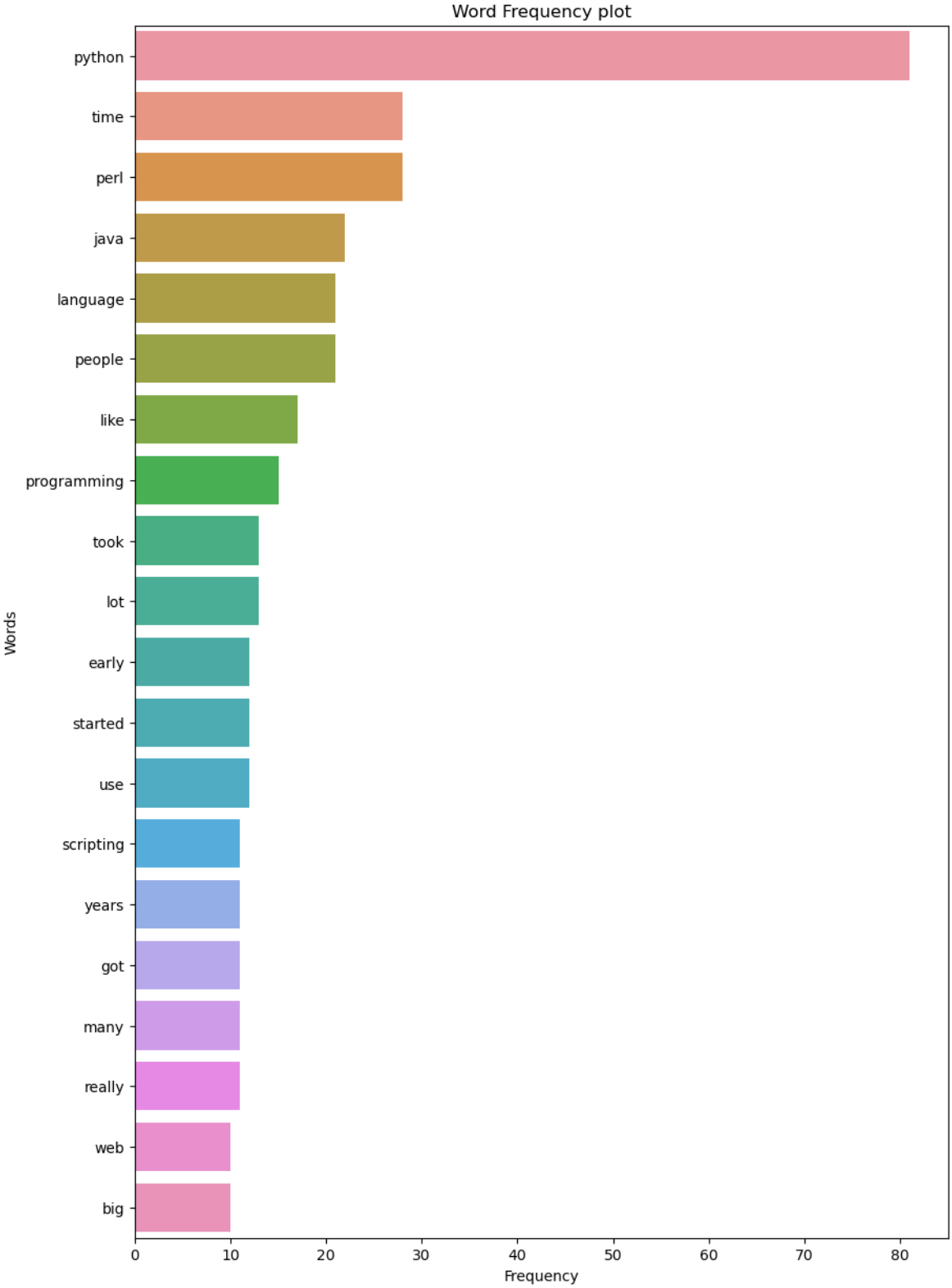
Most repeated words in comments

In [9]:
```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

sw = set(stopwords.words('english'))

comments = fp_comments['body'].values.tolist()

corpus = [word.lower() for comment in comments for word in word_tokenize(
corpus[:5]
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/slowgamer/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[9]: ['hardware', 'ready', 'python', 'time', 'python']

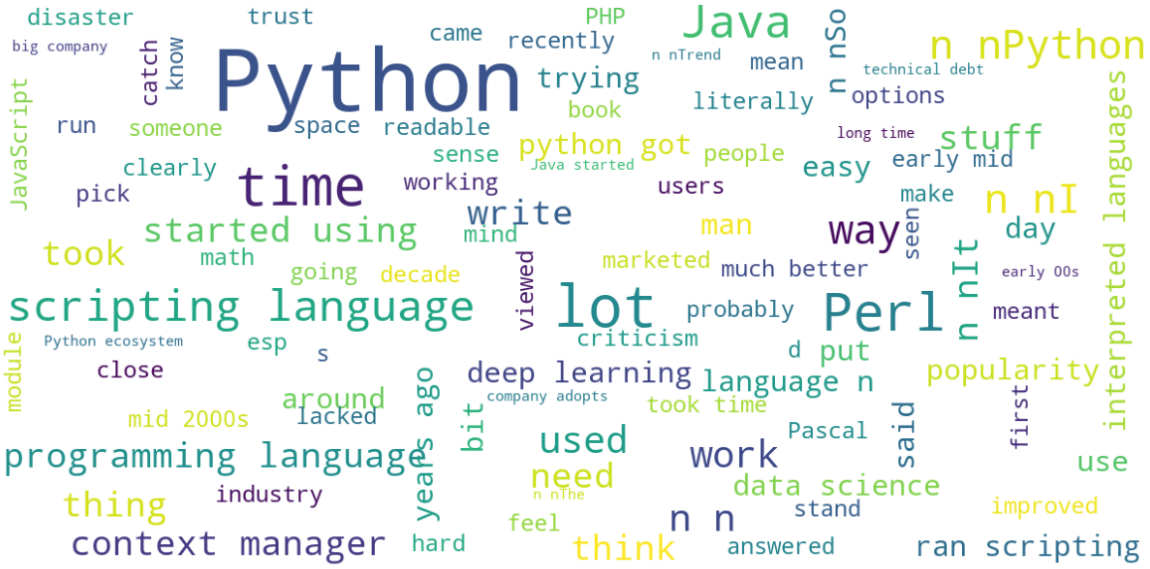
In [10]:
```
from collections import Counter

fig,ax = plt.subplots(1,1,figsize=(10,15))
word_counts = dict(Counter(corpus).most_common(20))
ax.set_xlabel('Frequency')
ax.set_ylabel('Words')
ax.set_title('Word Frequency plot')
sns.barplot(x=list(word_counts.values()),y=list(word_counts.keys()),ax=ax
plt.show()
```



In [52]:
```
from wordcloud import WordCloud, STOPWORDS

sw = set(STOPWORDS)

def generate_cloud(data,n_grams=2):
    fig, ax = plt.subplots(1,1,figsize=(15,15))
    wordcloud = WordCloud(
        background_color='white',
        stopwords=sw,
        max_words=100,
        max_font_size=30,
        scale=3,
        random_state=1,
        collocation_threshold=n_grams
    ).generate(data)

    ax.imshow(wordcloud)
    ax.axis('off')
    plt.show()
```

In [12]:
```
generate_cloud(str(fp_comments['body'].values).replace('\n',' '))
```

In [13]:
```
generate_cloud(str(fp_comments['body'].values).replace('\n',' '),n_grams=
```



Sentiment Analysis

Using Text Blob Library

In [16]:
```
from textblob import TextBlob

def polarity_score(text):
    return TextBlob(text).sentiment.polarity
```

In [30]:
```
fp_comments['comment_sent_score'] =  fp_comments['body'].apply(lambda tex
```

```
/tmp/ipykernel_30060/1826135865.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  fp_comments['comment_sent_score'] =  fp_comments['body'].apply(lambda te
xt: polarity_score(text))
```

In [31]:
```
fp_comments[['body','comment_sent_score']]
```