Instagram Analysis

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from sklearn.model_selection import train_test_split
from sklearn.linear_model import PassiveAggressiveRegressor

data = pd.read_csv("Instagram data.csv", encoding = 'latin1')
print(data.head())
```

```
   Impressions  From Home  From Hashtags  From Explore  From Other  Saves  \
0         3920       2586           1028           619          56     98
1         5394       2727           1838          1174          78    194
2         4021       2085           1188             0         533     41
3         4528       2700            621           932          73    172
4         2518       1704            255           279          37     96

   Comments  Shares  Likes  Profile Visits  Follows  \
0         9       5    162              35        2
1         7      14    224              48       10
2        11       1    131              62       12
3        10       7    213              23        8
4         5       4    123               8        0

                                             Caption  \
0  Here are some of the most important data visua...
1  Here are some of the best data science project...
2  Learn how to train a machine learning model an...
3  Here�s how you can write a Python program to d...
4  Plotting annotations while visualizing your da...

                                            Hashtags
0  #finance #money #business #investing #investme...
1  #healthcare #health #covid #data #datascience ...
2  #data #datascience #dataanalysis #dataanalytic...
3  #python #pythonprogramming #pythonprojects #py...
4  #datavisualization #datascience #data #dataana...
```

In [2]:

```
#Before starting everything, let’s have a look at whether this dataset contains
data.isnull().sum()
```

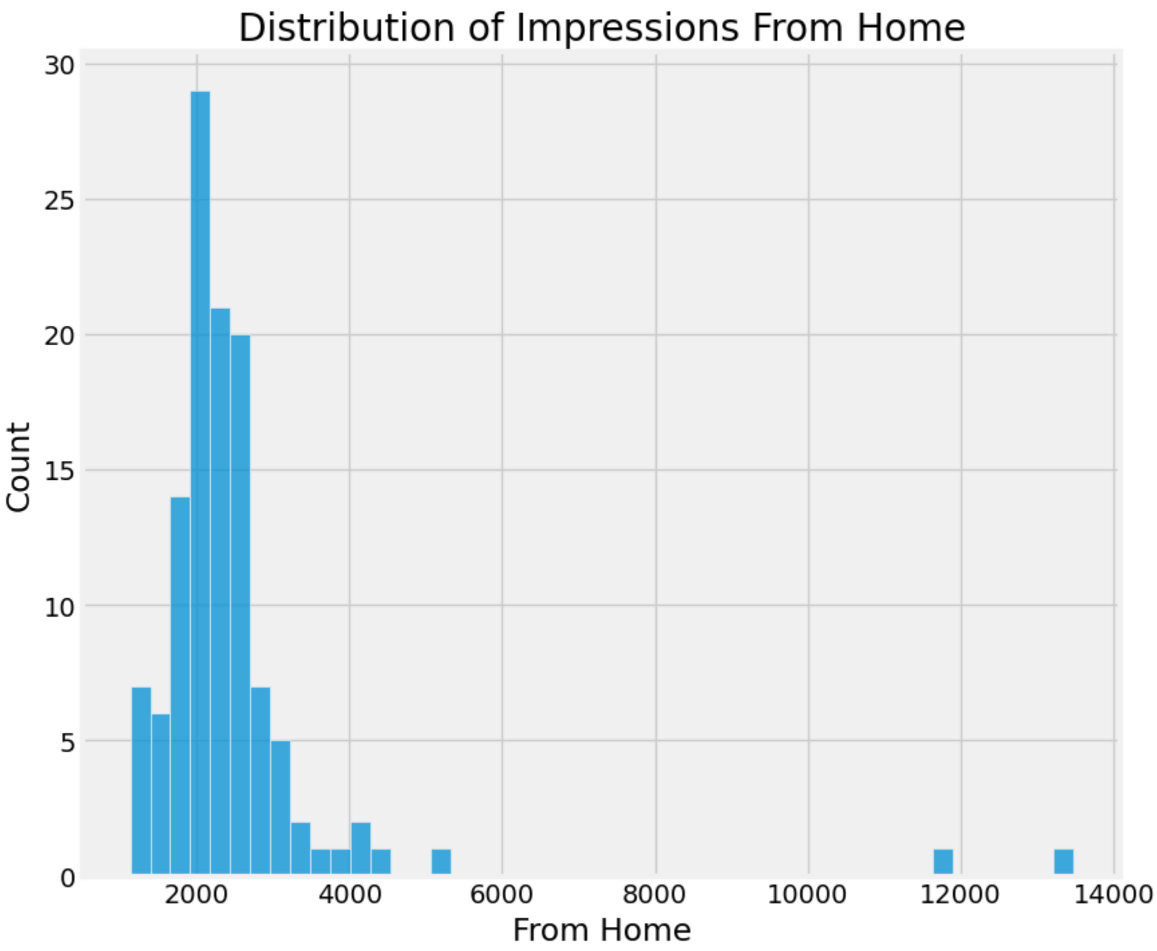
```
Out[2]: Impressions       0
        From Home         0
        From Hashtags     0
        From Explore      0
        From Other        0
        Saves             0
        Comments          0
        Shares            0
        Likes             0
        Profile Visits    0
        Follows           0
        Caption           0
        Hashtags          0
        dtype: int64
```

```
In [3]: #Let’s have a look at the insights of the columns to understand the data type of
        data.info()
```

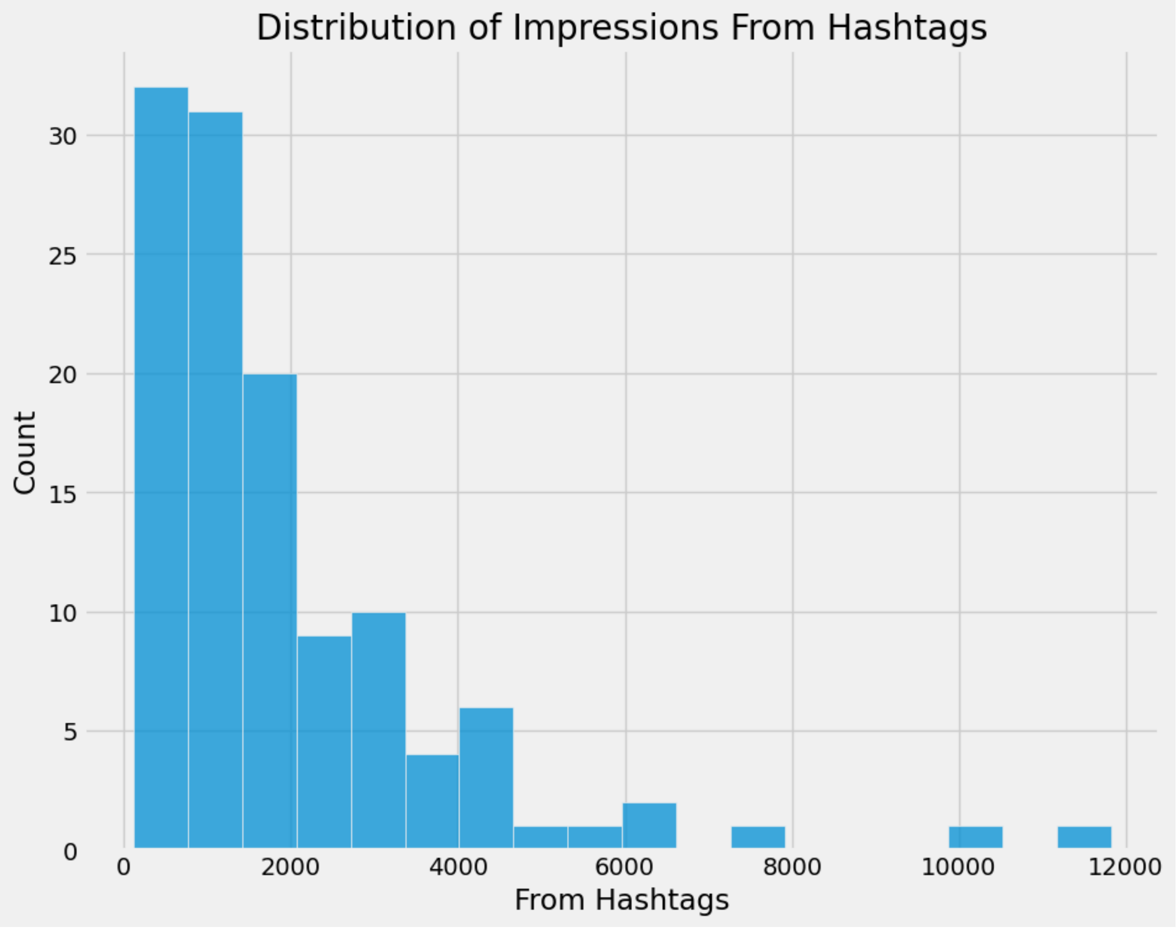
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119 entries, 0 to 118
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Impressions     119 non-null    int64 
 1   From Home       119 non-null    int64 
 2   From Hashtags   119 non-null    int64 
 3   From Explore    119 non-null    int64 
 4   From Other      119 non-null    int64 
 5   Saves           119 non-null    int64 
 6   Comments        119 non-null    int64 
 7   Shares          119 non-null    int64 
 8   Likes           119 non-null    int64 
 9   Profile Visits  119 non-null    int64 
 10  Follows         119 non-null    int64 
 11  Caption         119 non-null    object
 12  Hashtags        119 non-null    object
dtypes: int64(11), object(2)
memory usage: 12.2+ KB
```

Analyzing Instagram Reach

```
In [4]: #Analyzing the reach of my Instagram posts. I will first have a look at the dist
        plt.figure(figsize=(10, 8))
        plt.style.use('fivethirtyeight')
        plt.title("Distribution of Impressions From Home")
        sns.histplot(data['From Home'])
        plt.show()
```

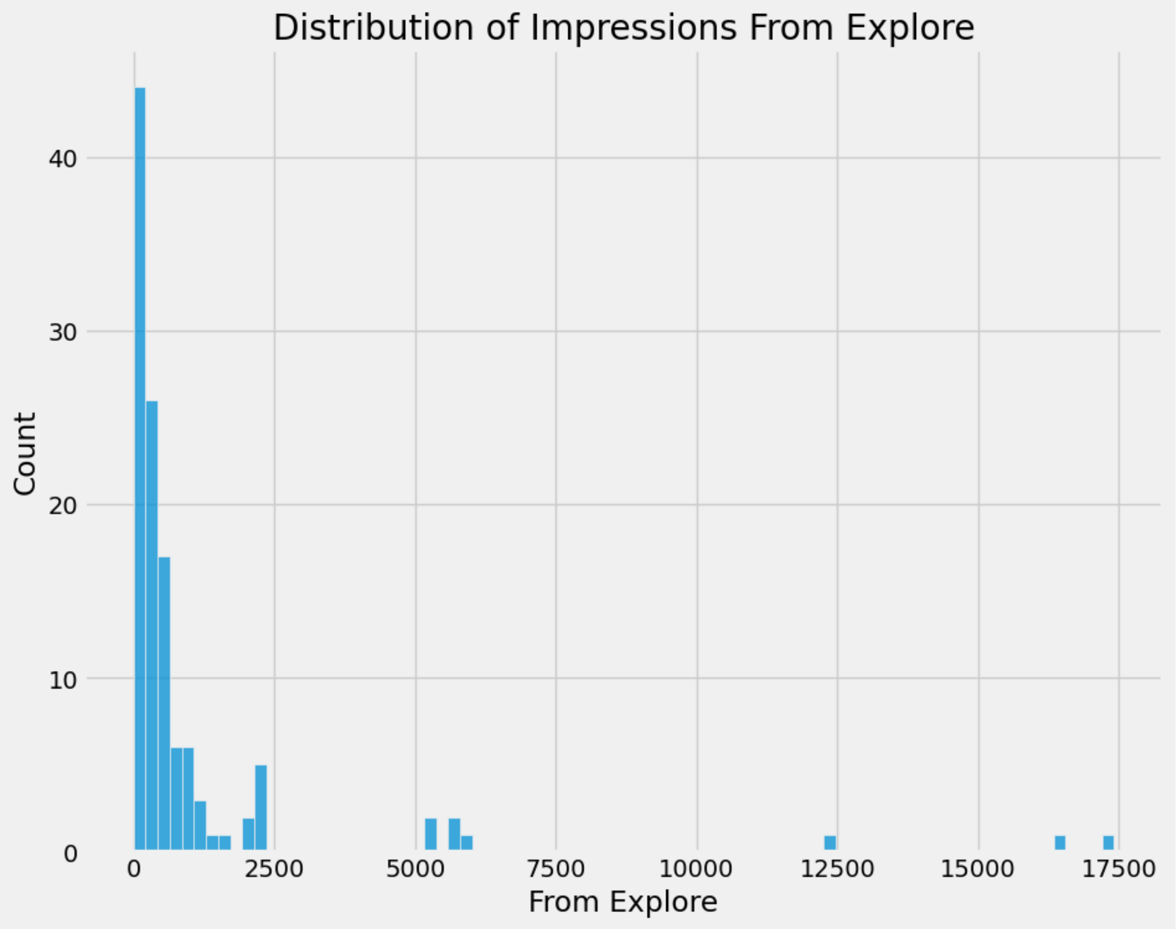


In [5]:
```
#The impressions I get from the home section on Instagram shows how much my post
plt.figure(figsize=(10, 8))
plt.title("Distribution of Impressions From Hashtags")
sns.histplot(data['From Hashtags'])
plt.show()
```



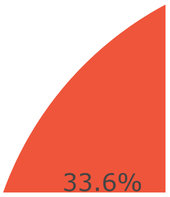
In [6]:

```
#Now let’s have a look at the distribution of impressions I have received from t
plt.figure(figsize=(10, 8))
plt.title("Distribution of Impressions From Explore")
sns.histplot(data['From Explore'])
plt.show()
```



In [7]:
```
#Now let’s have a look at the percentage of impressions I get from various sourc
home = data["From Home"].sum()
hashtags = data["From Hashtags"].sum()
explore = data["From Explore"].sum()
other = data["From Other"].sum()

labels = ['From Home','From Hashtags','From Explore','Other']
values = [home, hashtags, explore, other]

fig = px.pie(data, values=values, names=labels,
             title='Impressions on Instagram Posts From Various Sources', hole=0
fig.show()
```

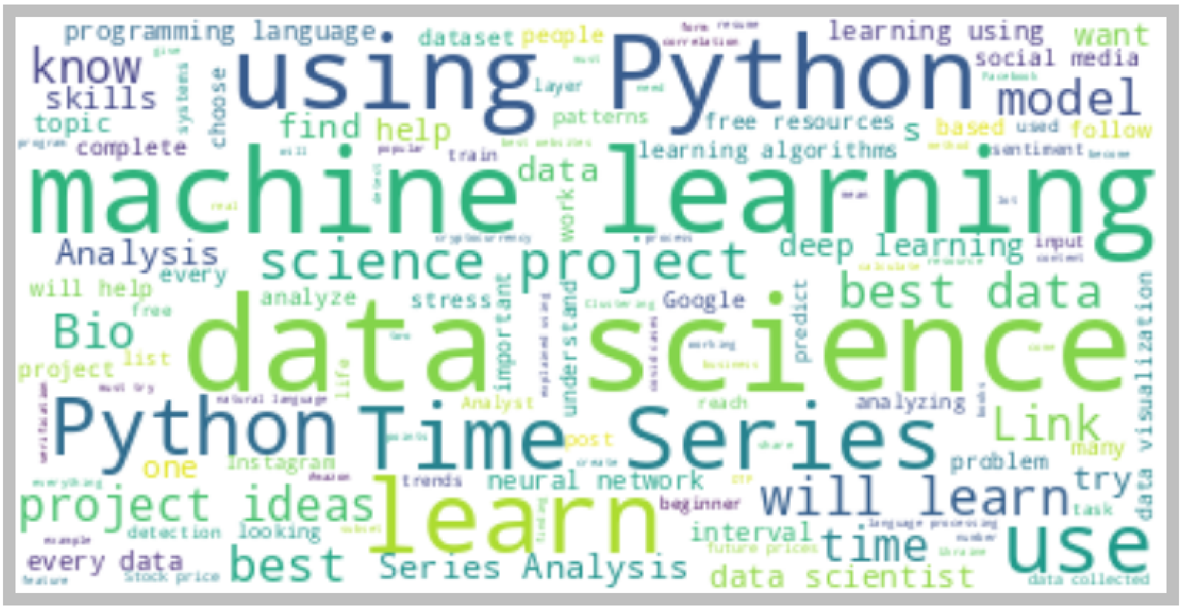
Impressions on Instagram Posts From Various Sources



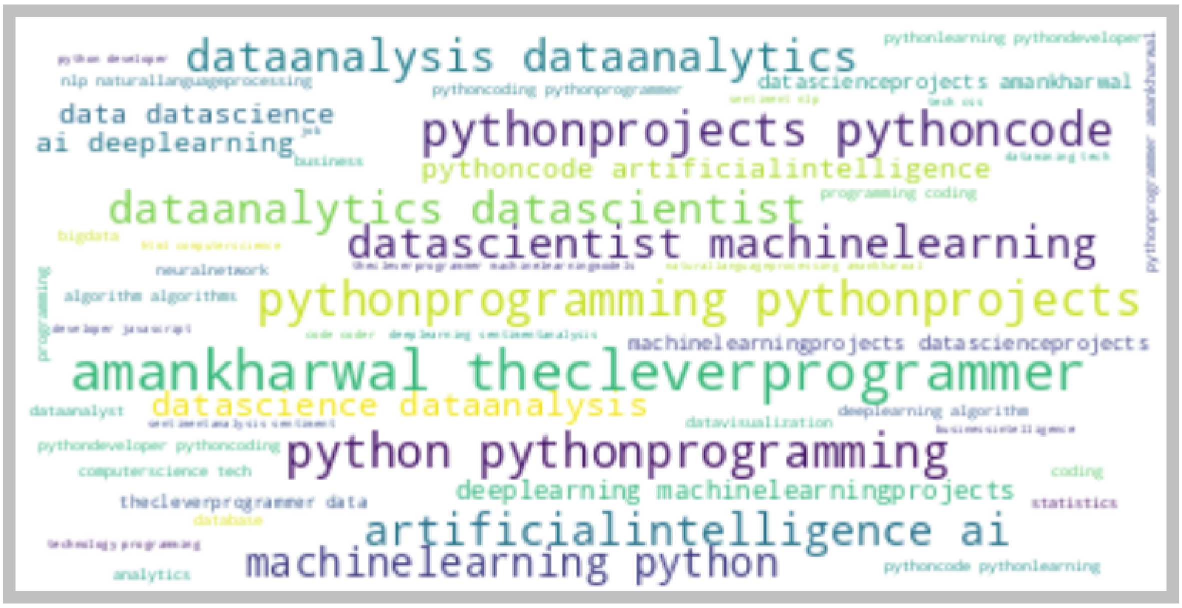
So the above donut plot shows that almost 50 per cent of the reach is from my followers, 38.1 per cent is from hashtags, 9.14 per cent is from the explore section, and 3.01 per cent is from other sources.

Analyzing Content

In [8]:
```
#Now let’s analyze the content of my Instagram posts.Let’s create a wordcloud of
text = " ".join(i for i in data.Caption)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(te
plt.style.use('classic')
plt.figure( figsize=(12,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



In [9]:
```
#Now let’s create a wordcloud of the hashtags column to look at the most used ha
text = " ".join(i for i in data.Hashtags)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(te
plt.figure( figsize=(12,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



In [10]:
```
pip install statsmodels
```

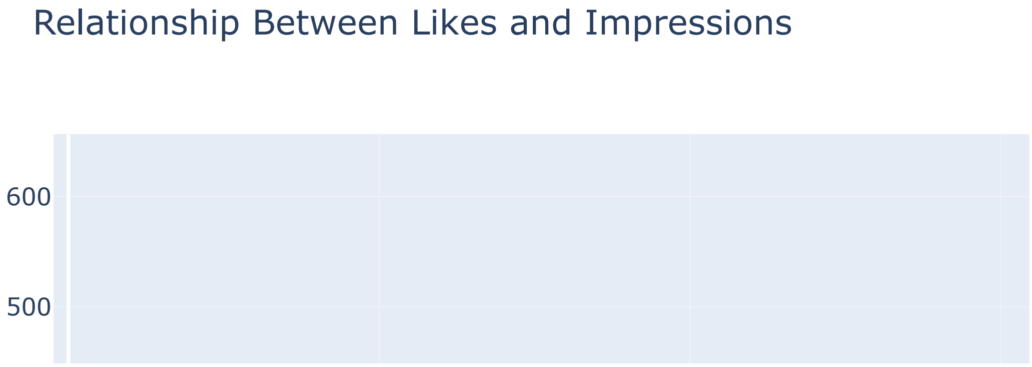
```
Requirement already satisfied: statsmodels in c:\users\sethu\appdata\local\progra
ms\python\python311\lib\site-packages (0.14.0)Note: you may need to restart the k
ernel to use updated packages.

Requirement already satisfied: numpy>=1.18 in c:\users\sethu\appdata\local\progra
ms\python\python311\lib\site-packages (from statsmodels) (1.25.0)
Requirement already satisfied: scipy!=1.9.2,>=1.4 in c:\users\sethu\appdata\local
\programs\python\python311\lib\site-packages (from statsmodels) (1.11.1)
Requirement already satisfied: pandas>=1.0 in c:\users\sethu\appdata\local\progra
ms\python\python311\lib\site-packages (from statsmodels) (2.0.3)
Requirement already satisfied: patsy>=0.5.2 in c:\users\sethu\appdata\local\progr
ams\python\python311\lib\site-packages (from statsmodels) (0.5.3)
Requirement already satisfied: packaging>=21.3 in c:\users\sethu\appdata\local\pr
ograms\python\python311\lib\site-packages (from statsmodels) (23.1)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\sethu\appdata\l
ocal\programs\python\python311\lib\site-packages (from pandas>=1.0->statsmodels)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\sethu\appdata\local\progr
ams\python\python311\lib\site-packages (from pandas>=1.0->statsmodels) (2023.3)
Requirement already satisfied: tzdata>=2022.1 in c:\users\sethu\appdata\local\pro
grams\python\python311\lib\site-packages (from pandas>=1.0->statsmodels) (2023.3)
Requirement already satisfied: six in c:\users\sethu\appdata\local\programs\pytho
n\python311\lib\site-packages (from patsy>=0.5.2->statsmodels) (1.16.0)
```

Analyzing Relationships

Now let’s analyze relationships to find the most important factors of our Instagram reach. It will also help us in understanding how the Instagram algorithm works.

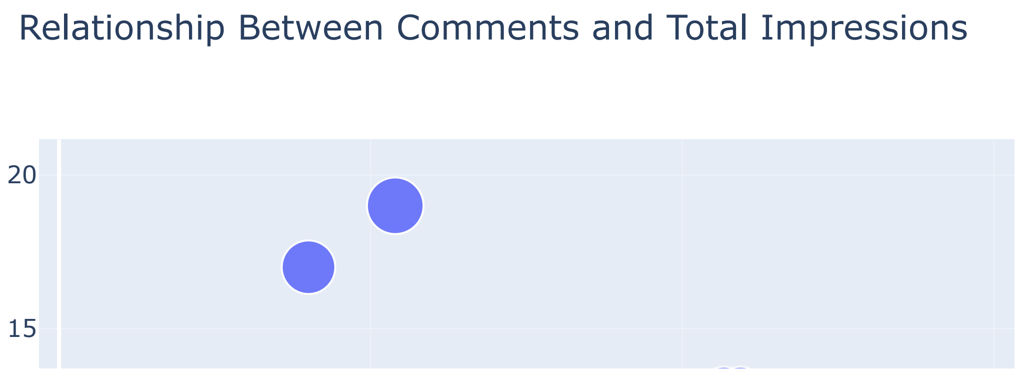
In [11]:

```
#Let’s have a look at the relationship between the number of likes and the numbe
figure = px.scatter(data_frame = data, x="Impressions",
                    y="Likes", size="Likes", trendline="ols",
                    title = "Relationship Between Likes and Impressions")
figure.show()
```



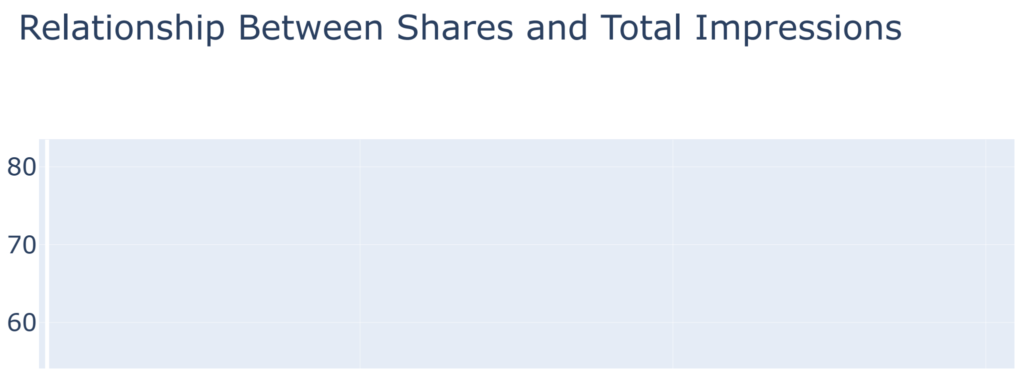
There is a linear relationship between the number of likes and the reach I got on Instagram.

In [12]:
```
#Now let’s see the relationship between the number of comments and the number of

figure = px.scatter(data_frame = data, x="Impressions",
                    y="Comments", size="Comments", trendline="ols",
                    title = "Relationship Between Comments and Total Impressions
figure.show()
```



It looks like the number of comments we get on a post doesn’t affect its reach.

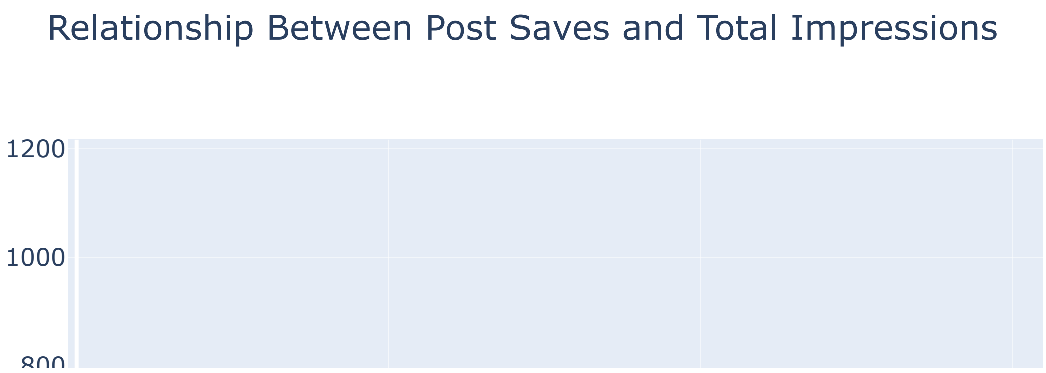
In [13]:
```
 #Now let’s have a look at the relationship between the number of shares and the
figure = px.scatter(data_frame = data, x="Impressions",
                    y="Shares", size="Shares", trendline="ols",
                    title = "Relationship Between Shares and Total Impressions")
figure.show()
```



A more number of shares will result in a higher reach, but shares don’t affect the reach of a post as much as likes do.

In [14]:

```
#Now let’s have a look at the relationship between the number of saves and the n
figure = px.scatter(data_frame = data, x="Impressions",
                    y="Saves", size="Saves", trendline="ols",
                    title = "Relationship Between Post Saves and Total Impressio
figure.show()
```



There is a linear relationship between the number of times my post is saved and the reach of my Instagram post.

In [15]:
```
#Now let’s have a look at the correlation of all the columns with the Impression
#Exclude non-numeric columns from correlation calculation
numeric_data = data.select_dtypes(include='number')
correlation = numeric_data.corr()
print(correlation["Impressions"].sort_values(ascending=False))
```

```
Impressions       1.000000
From Explore      0.893607
Follows           0.889363
Likes             0.849835
From Home         0.844698
Saves             0.779231
Profile Visits    0.760981
Shares            0.634675
From Other        0.592960
From Hashtags     0.560760
Comments         -0.028524
Name: Impressions, dtype: float64
```

So we can say that more likes and saves will help you get more reach on Instagram.The higher number of shares will also help you get more reach, but a low number of shares will not affect your reach either.

Analyzing Conversion Rate

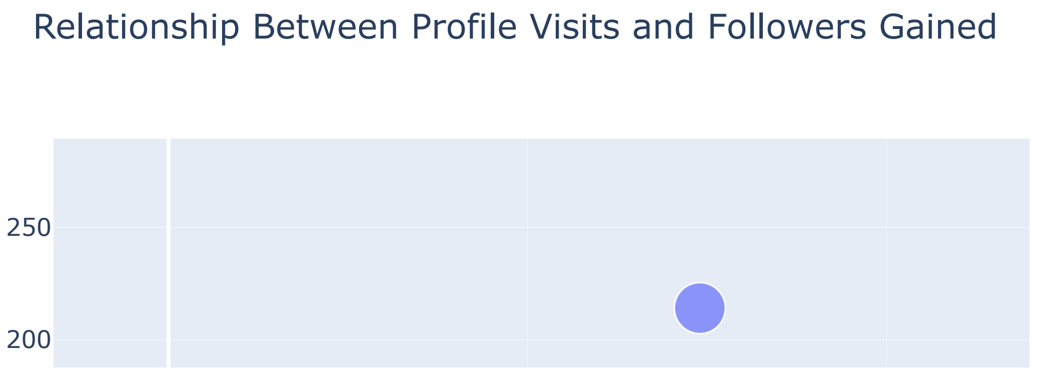
In Instagram, conversation rate means how many followers you are getting from the number of profile visits from a post.The formula that you can use to calculate conversion rate is (Follows/Profile Visits) * 100.

```
In [16]: #Now let’s have a look at the conversation rate of my Instagram account:
         conversion_rate = (data["Follows"].sum() / data["Profile Visits"].sum()) * 100
         print(conversion_rate)
```

```
41.00265604249668
```

So the conversation rate of my Instagram account is 41% which sounds like a very good conversation rate.

```
In [17]: #Let’s have a look at the relationship between the total profile visits and the
         figure = px.scatter(data_frame = data, x="Profile Visits",
                             y="Follows", size="Follows", trendline="ols",
                             title = "Relationship Between Profile Visits and Followers G
         figure.show()
```



The relationship between profile visits and followers gained is also linear.

Instagram Reach Prediction Model

In [18]:

```
#I will train a machine learning model to predict the reach of an Instagram post
x = np.array(data[['Likes', 'Saves', 'Comments', 'Shares',
                   'Profile Visits', 'Follows']])
y = np.array(data["Impressions"])
xtrain, xtest, ytrain, ytest = train_test_split(x, y,
                                                test_size=0.2,
                                                random_state=42)
```

In [19]:

```
#Now here’s is how we can train a machine learning model to predict the reach of
model = PassiveAggressiveRegressor()
model.fit(xtrain, ytrain)
model.score(xtest, ytest)
```

Out[19]:

```
0.8806402989874751
```

In [20]:

```
#Now let’s predict the reach of an Instagram post by giving inputs to the machin
# Features = [['Likes','Saves', 'Comments', 'Shares', 'Profile Visits', 'Follows
features = np.array([[282.0, 233.0, 4.0, 9.0, 165.0, 54.0]])
model.predict(features)
```

Out[20]:

```
array([11377.19475193])
```

So this is how you can analyze and predict the reach of Instagram posts with machine learning using Python. If a content creator wants to do well on Instagram in a long run, they have to look at the data of their Instagram reach.