Data preprocessing

In [25]:
```
# Importing neccessary packages
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.pyplot import pie, axis, show
from pandas import DataFrame
```

In [11]:
```
# Loading the dataset
tips = sns.load_dataset("tips")
```

In [12]:
```
# Checking the structure of tips dataset.
type(tips)
```

Out[12]:
```
pandas.core.frame.DataFrame
```

In [13]:
```
# Having an overview of
tips.head()
```

Out[13]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

In [14]:
```
# Getting information about the data
tips.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype   
---  ------      --------------  -----   
 0   total_bill  244 non-null    float64 
 1   tip         244 non-null    float64 
 2   sex         244 non-null    category
 3   smoker      244 non-null    category
 4   day         244 non-null    category
 5   time        244 non-null    category
 6   size        244 non-null    int64   
dtypes: category(4), float64(2), int64(1)
memory usage: 7.4 KB
```

Data Modelling

> I will be creating the following columns for the purpose of analysis:

- tip_ratio - ratio of tip to bill,

- sum - the sum of total bill and tip
- percent - tip as a percent of sum,
- bpp - amount of total bill divided by number of people in the group,
- tpp - amount of tip per person.

In [15]:
```
# new column created - percent of tip
tips["tip_ratio"] = tips["tip"]/tips["total_bill"]

# new column created - sum of total bill and tip
tips["sum"] = tips["total_bill"]+tips["tip"] # appended at the end of the array

# new column created - ratio of tip to sum
tips["percent"] = round(tips["tip"]/tips["sum"]*100, 2)

# add column: bpp - bill per person
tips["bpp"] = tips["total_bill"]/tips["size"]

# add column: tpp - tip per person
tips["tpp"] = tips["tip"]/tips["size"]
```

In [16]:
```
# Checking if the columns were sucessfully added
tips.head()
```

Out[16]:

	total_bill	tip	sex	smoker	day	time	size	tip_ratio	sum	percent	bpp	tpp
0	16.99	1.01	Female	No	Sun	Dinner	2	0.059447	18.00	5.61	8.495000	0.505000
1	10.34	1.66	Male	No	Sun	Dinner	3	0.160542	12.00	13.83	3.446667	0.553333
2	21.01	3.50	Male	No	Sun	Dinner	3	0.166587	24.51	14.28	7.003333	1.166667
3	23.68	3.31	Male	No	Sun	Dinner	2	0.139780	26.99	12.26	11.840000	1.655000
4	24.59	3.61	Female	No	Sun	Dinner	4	0.146808	28.20	12.80	6.147500	0.902500

Data analysis and exploration

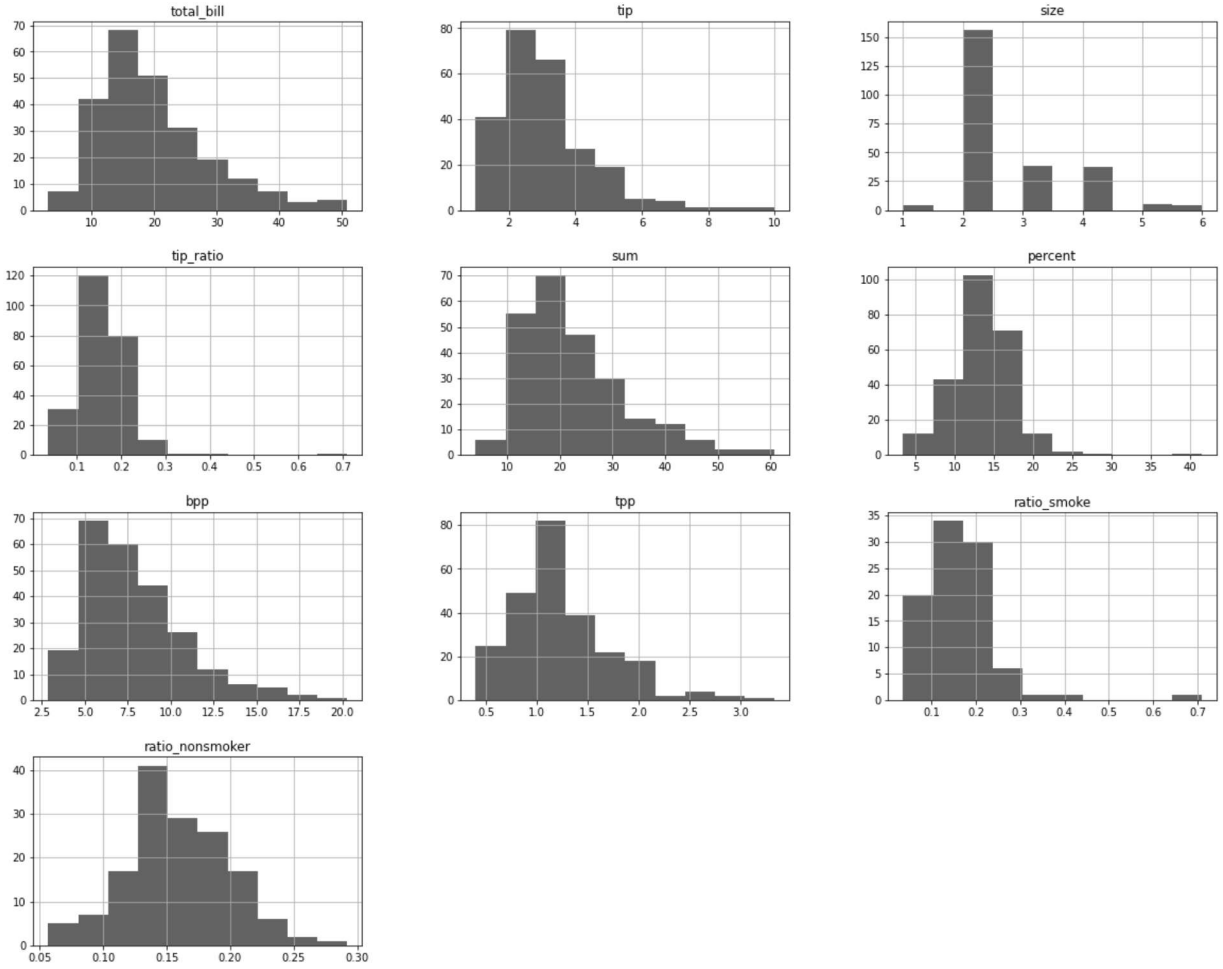
In [17]:
```
# Getting the numerical summary of the data.
tips.describe()
```

Out[17]:

	total_bill	tip	size	tip_ratio	sum	percent	bpp	tpp
count	244.000000	244.000000	244.000000	244.000000	244.000000	244.000000	244.00000	244.00000(
mean	19.785943	2.998279	2.569672	0.160803	22.784221	13.643320	7.88823	1.21276
std	8.902412	1.383638	0.951100	0.061072	9.890116	4.058544	2.91435	0.49170
min	3.070000	1.000000	1.000000	0.035638	4.070000	3.440000	2.87500	0.40000(
25%	13.347500	2.000000	2.000000	0.129127	15.475000	11.437500	5.80250	0.86250(
50%	17.795000	2.900000	2.000000	0.154770	20.600000	13.405000	7.25500	1.10750(
75%	24.127500	3.562500	3.000000	0.191475	27.722500	16.070000	9.39000	1.50000(
max	50.810000	10.000000	6.000000	0.710345	60.810000	41.530000	20.27500	3.33333

In [35]:
```
## Getting the graphical summary of the data.
tips.hist(figsize=(20,16))
```

Out[35]:
```
array([[<AxesSubplot:title={'center':'total_bill'}>,
        <AxesSubplot:title={'center':'tip'}>,
        <AxesSubplot:title={'center':'size'}>],
       [<AxesSubplot:title={'center':'tip_ratio'}>,
        <AxesSubplot:title={'center':'sum'}>,
        <AxesSubplot:title={'center':'percent'}>],
       [<AxesSubplot:title={'center':'bpp'}>,
        <AxesSubplot:title={'center':'tpp'}>,
        <AxesSubplot:title={'center':'ratio_smoke'}>],
       [<AxesSubplot:title={'center':'ratio_nonsmoker'}>, <AxesSubplot:>,
        <AxesSubplot:>]], dtype=object)
```



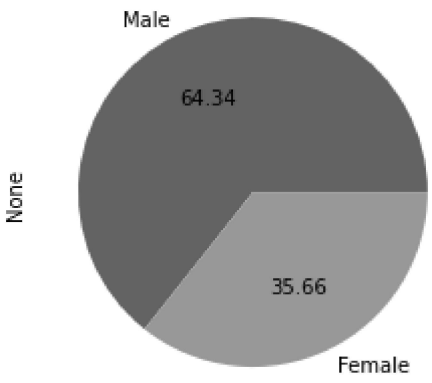
- The highest amount customers paid(excluding the tips) in the resturant was 50.81 Euros and the minimum amount was 3.07 Euros.
- The maximum amount of tips paid by customer was 10 Euros while the minimum tip paid was 1 euros.
- On average, the percent of tip is 13.64%.
- On average, total bill split equally in the group is 7.89.
- On average, tip per person is 1.21.

In [18]:
```
# Getting the proportion of gender
print(tips.sex.value_counts())
```

```
Male      157
Female     87
Name: sex, dtype: int64
```

In [19]:
```
tips.groupby('sex').size().plot(kind='pie', autopct='%.2f')
```

Out[19]:
```
<AxesSubplot:ylabel='None'>
```



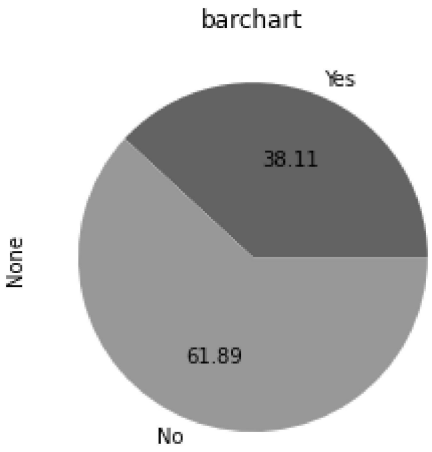
> The resturant has more male customers than the females, with 64.34% and 35.66% respectively.

In [20]:
```
# Getting the unique values that's in the smoker column
print(tips.smoker.value_counts())
```

```
No     151
Yes     93
Name: smoker, dtype: int64
```

In [21]:
```
tips.groupby('smoker').size().plot(kind='pie', autopct='%.2f', title = "barchart")
```

Out[21]:
```
<AxesSubplot:title={'center':'barchart'}, ylabel='None'>
```



> Non-smokers patronizes the resturants more than the smokers with a percetange of 61.89 and 38.11 respectively.

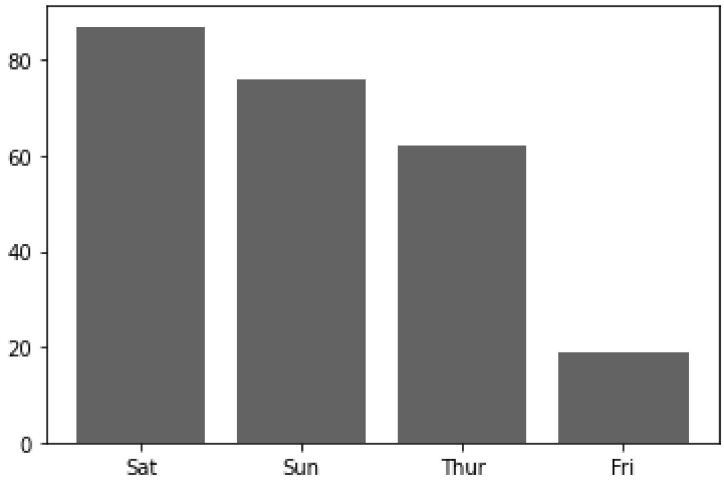
In [22]:
```
print("Unique values in column 'day':\t", sorted(tips.day.unique()))
print("Unique values in column 'time':\t", sorted(tips.time.unique()))
```

```
Unique values in column 'day':   ['Fri', 'Sat', 'Sun', 'Thur']
Unique values in column 'time':  ['Dinner', 'Lunch']
```

> It appears the resutarnt only opens on Thursday, Friday, Saturday and Sunday and offers their services at either lunch or dinner time.

In [26]:
```
# Plotting a bar chart to determine the day which tips are mostly paid
plt.bar(tips['day'].value_counts().index,tips['day'].value_counts().values)
```

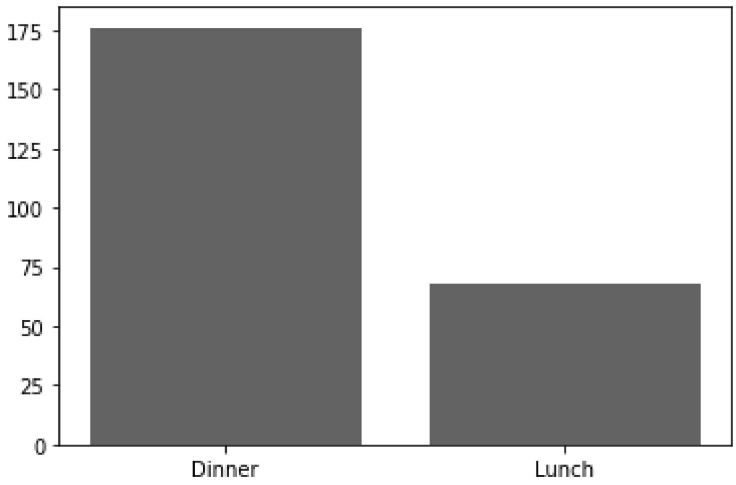
Out[26]:
```
<BarContainer object of 4 artists>
```



> Tips are mostly paid on Saturdays.

In [27]:
```
# Plotting a bar chart to dermine which period was tip mostly paid
plt.bar(tips['time'].value_counts().index,tips['time'].value_counts().values)
```

Out[27]:
```
<BarContainer object of 2 artists>
```



> Tips was mostly paid during the dinner period

In [28]:
```
### I want to compare the average of tip per smoker and non-smoker

# mean tip per smoking person
print("Average tip per smoker:\t\t\t", tips[tips["smoker"] == "Yes"]["tpp"].mean())

# mean total bill among smokers
print("Average total bill among smokers:\t\t", tips[tips["smoker"] == "Yes"]["total_
```

```
# Tip ratio among smokers
tips["ratio_smoke"] = tips[tips["smoker"] == "Yes"]["tip"] / tips[tips["smoker"] ==
print("Ratio of tips to total bill among smokers:\t", tips["ratio_smoke"].mean())


# mean tip per smoking person
print("Average tip per non-smoker:\t\t\t", tips[tips["smoker"] == "No"]["tpp"].mean(

# mean total bill among smokers
print("Average total bill among non-smokers:\t\t", tips[tips["smoker"] == "No"]["tot

# mean ratio among smokers
tips["ratio_nonsmoker"] = tips[tips["smoker"] == "No"]["tip"] / tips[tips["smoker"]
print("Ratio of tips to total bill among non- smokers:\t", tips["ratio_nonsmoker"].m
```

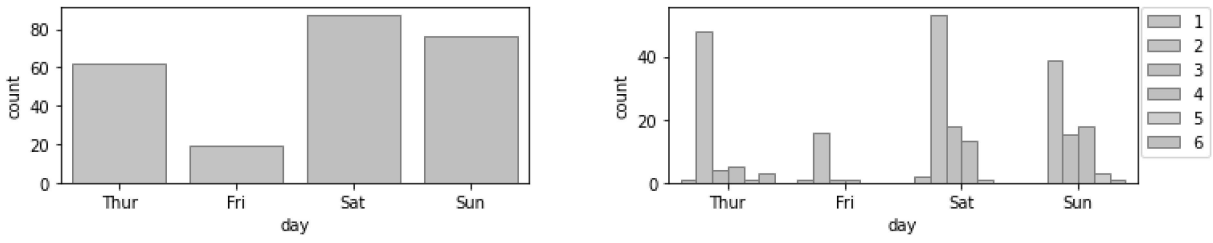
```
Average tip per smoker:                  1.2977956989247312
Average total bill among smokers:                20.756344086021507
Ratio of tips to total bill among smokers:       0.1631960446368779
Average tip per non-smoker:                      1.1603896247240624
Average total bill among non-smokers:            19.18827814569537
Ratio of tips to total bill among non- smokers:  0.1593284621792153
```

> The smokers pay more for the service as well as give higher tips. On average the ratio tip to total bill among smokers is approximately 16.32%

In [29]:

```
# Creating plots to show the number of customers split into days the restaurant was

# Display setting
fig, ax = plt.subplots(ncols=2, figsize=(12, 2))
plt.suptitle("Figure 2. Number of clients on each day.", y=1.1, fontsize=14)
plt.subplots_adjust(left=None, bottom=None, right=None, top=None, wspace=.3, hspace=

# plot data and propertiesy
sns.countplot(data=tips, x="day", palette="pastel", edgecolor=".5", ax=ax[0]) # numb
sns.countplot(data=tips, x="day", palette="pastel", edgecolor=".5", hue='size', ax=a
plt.legend(bbox_to_anchor=(1.01, 1), loc=2, borderaxespad=0.) # Put the legend out o

plt.show()
```

Figure 2. Number of clients on each day.

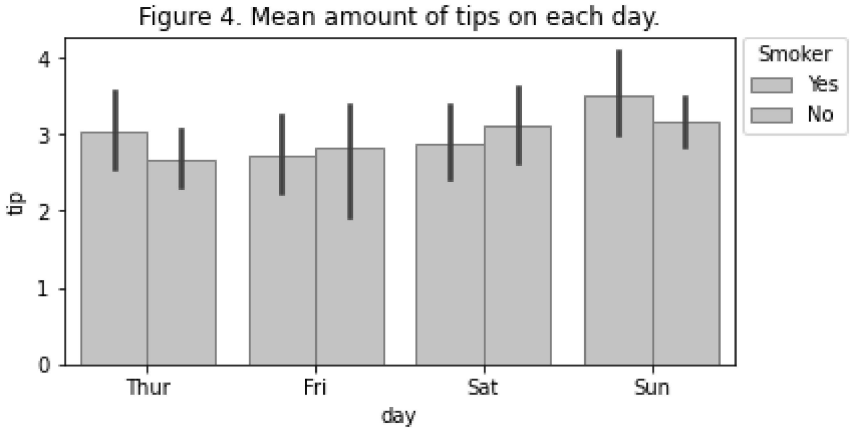


> There are more cutomers on Saturdays and the party size is mostly 2 accross all days

In [30]:

```
# Creating a bar plot that displays the amount of tip offered each day, categorized

# adapted from https://medium.com/@neuralnets/data-visualization-with-python-and-sea
fig, ax = plt.subplots(ncols=1, figsize=(6,3))
plt.title('Figure 4. Mean amount of tips on each day.')

# Mean amount of tip on each day
sns.barplot(data = tips, x = "day", y = "tip", hue="smoker", palette="pastel", edgec
```

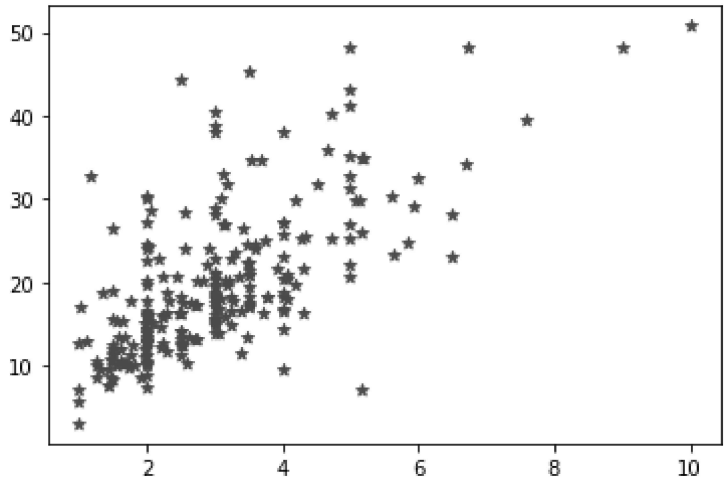
```
plt.legend(bbox_to_anchor=(1.01, 1), loc=2, borderaxespad=0.0, title="Smoker") # leg

plt.show()
```



> The above barplot is also affirming that the non-smokers pay more tips than the smokers.

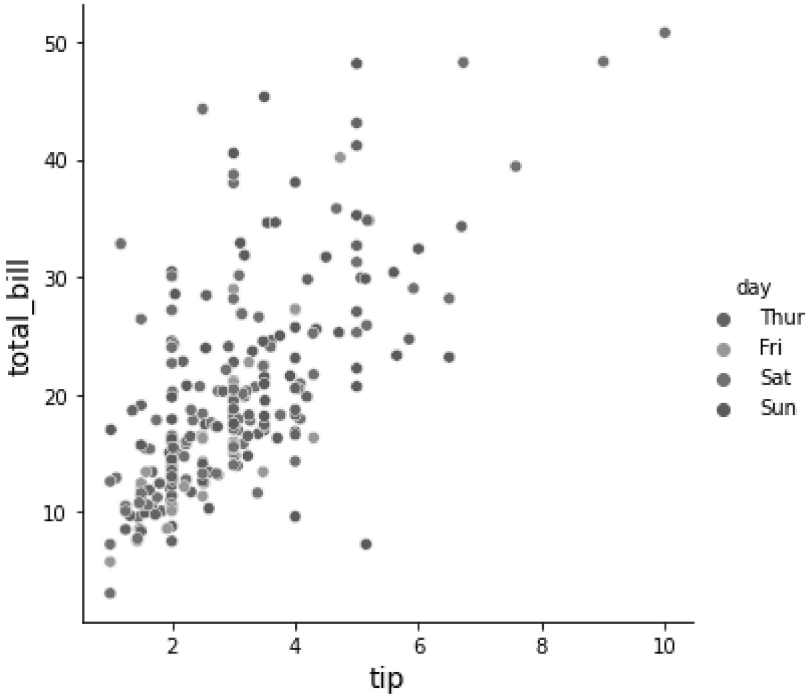
In [31]:
```
# Creating a scatter to see if there is relaltionship between tips and total bills p
plt.figure()
plt.scatter(tips.tip,tips.total_bill,color='g',marker='*')
```

Out[31]:
```
<matplotlib.collections.PathCollection at 0x216d0117bb0>
```



> There seem to be a positive linear relationship between tips and total bills.

In [32]:
```
# Creating a scattter plot to determine between bills and total_bills using days as
sns.relplot(x="tip", y="total_bill", data=tips, hue='day')
plt.xlabel('tip',fontsize=14)
plt.ylabel('total_bill',fontsize=14);
```



> Days of opening does not affect the positive relatinshop between tips and total bills. From the plot, it can also be seen that the highest tips was paid on a Saturday.

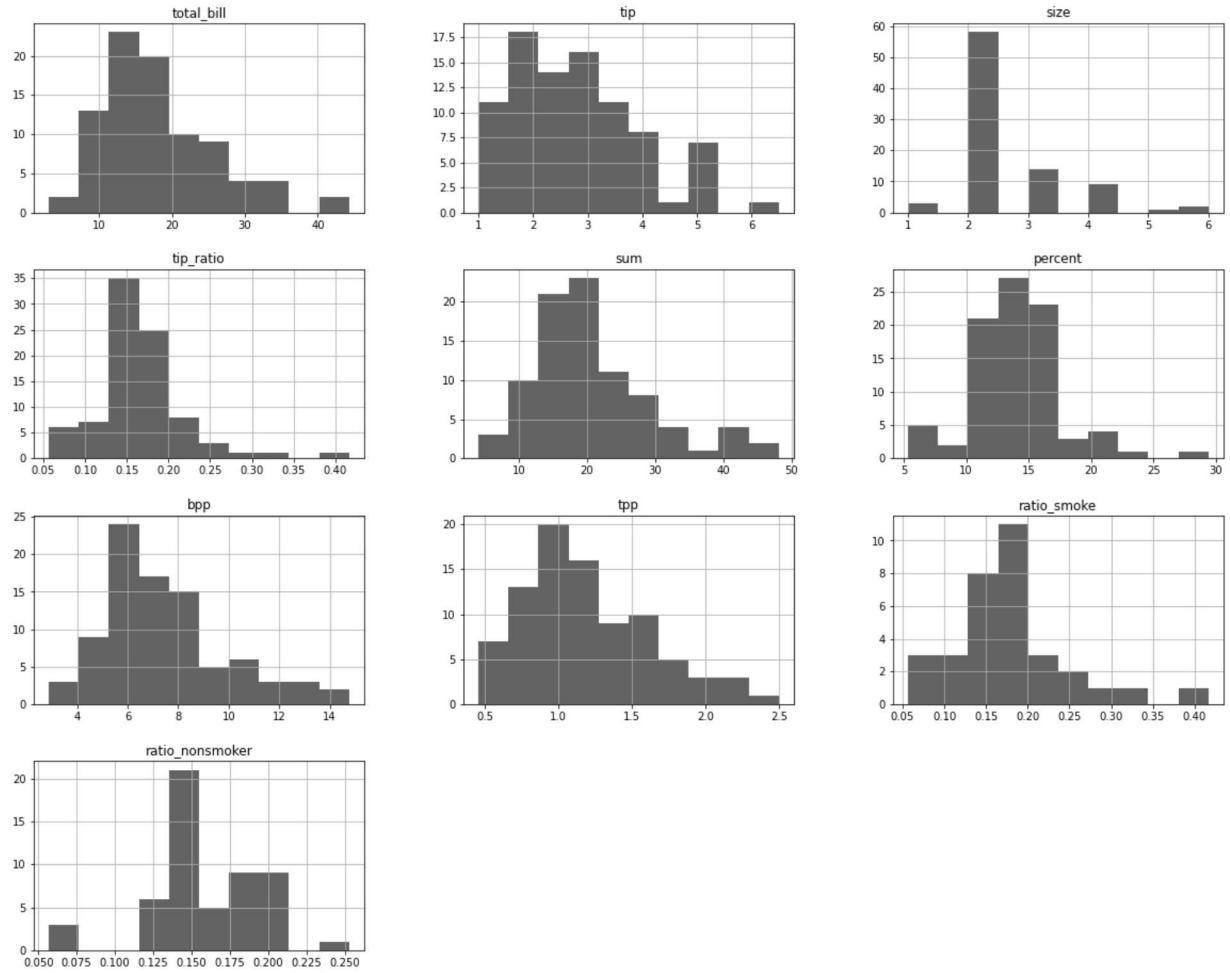
In [36]:
```
# Subsetting the data to have only female customers
Female_tips = tips.loc[tips.sex == 'Female']
Female_tips.head()
```

Out[36]:

	total_bill	tip	sex	smoker	day	time	size	tip_ratio	sum	percent	bpp	tpp
0	16.99	1.01	Female	No	Sun	Dinner	2	0.059447	18.00	5.61	8.495000	0.505000
4	24.59	3.61	Female	No	Sun	Dinner	4	0.146808	28.20	12.80	6.147500	0.902500
11	35.26	5.00	Female	No	Sun	Dinner	4	0.141804	40.26	12.42	8.815000	1.250000
14	14.83	3.02	Female	No	Sun	Dinner	2	0.203641	17.85	16.92	7.415000	1.510000
16	10.33	1.67	Female	No	Sun	Dinner	3	0.161665	12.00	13.92	3.443333	0.556667

In [37]:
```
# Graphical summary of the female customers
Female_tips.hist(figsize=(20,16))
```

Out[37]:
```
array([[<AxesSubplot:title={'center':'total_bill'}>,
        <AxesSubplot:title={'center':'tip'}>,
        <AxesSubplot:title={'center':'size'}>],
       [<AxesSubplot:title={'center':'tip_ratio'}>,
        <AxesSubplot:title={'center':'sum'}>,
        <AxesSubplot:title={'center':'percent'}>],
       [<AxesSubplot:title={'center':'bpp'}>,
        <AxesSubplot:title={'center':'tpp'}>,
        <AxesSubplot:title={'center':'ratio_smoke'}>],
       [<AxesSubplot:title={'center':'ratio_nonsmoker'}>, <AxesSubplot:>,
        <AxesSubplot:>]], dtype=object)
```



In [39]:
```
# Numerical summary of the female customers
Female_tips.describe()
```

Out[39]:

	total_bill	tip	size	tip_ratio	sum	percent	bpp	tpp	ratio_
count	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	33.
mean	18.056897	2.833448	2.459770	0.166491	20.890345	14.101724	7.463918	1.194603	0.
std	8.009209	1.159495	0.937644	0.053632	8.841798	3.781743	2.461519	0.428018	0.
min	3.070000	1.000000	1.000000	0.056433	4.070000	5.340000	2.875000	0.453333	0.
25%	12.750000	2.000000	2.000000	0.140416	14.870000	12.315000	5.682500	0.920000	0.
50%	16.400000	2.750000	2.000000	0.155581	18.900000	13.460000	6.710000	1.115000	0.
75%	21.520000	3.500000	3.000000	0.194266	25.000000	16.265000	8.721250	1.470000	0.
max	44.300000	6.500000	6.000000	0.416667	48.110000	29.410000	14.766667	2.500000	0.

For the females:

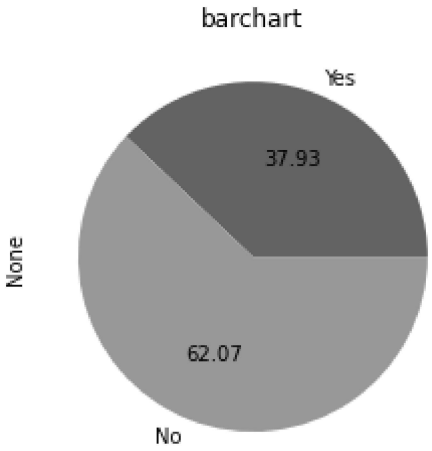
- The highest amount customers paid(excluding the tips) in the resturant was 44.3 Euros and the minimum amount was 3.07 Euros.
- The maximum amount of tips paid by customer was Euros while 6.5 the minimum tip paid was 1 euros.

In [36]:
```
# Getting the unique values for female smokers and non-smokers
print(female_tips.smoker.value_counts())
```

```
No     54
Yes    33
Name: smoker, dtype: int64
```

In [37]:
```
# Graphical representation of the unique values
female_tips.groupby('smoker').size().plot(kind='pie', autopct='%.2f', title = "barch
```

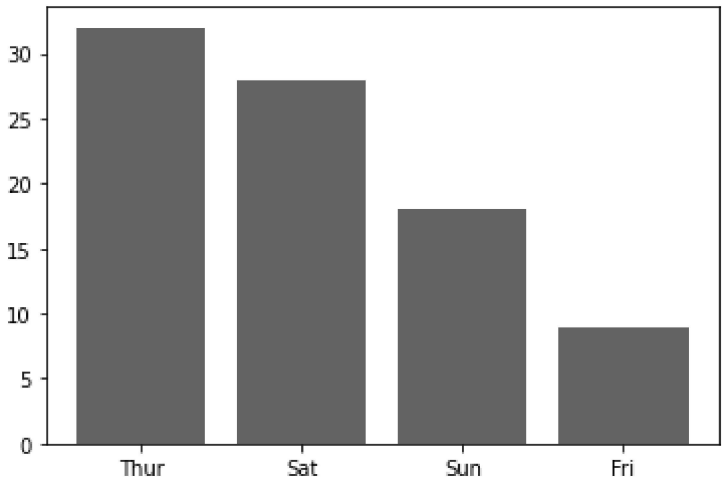
Out[37]:
```
<AxesSubplot:title={'center':'barchart'}, ylabel='None'>
```



> 62.07% of the females are non-smokers, while 37.93% are smokers

In [48]:
```
# Craeting a bar chart to determine which day did the females mostly paid tips
plt.bar(female_tips['day'].value_counts().index,female_tips['day'].value_counts().va
```

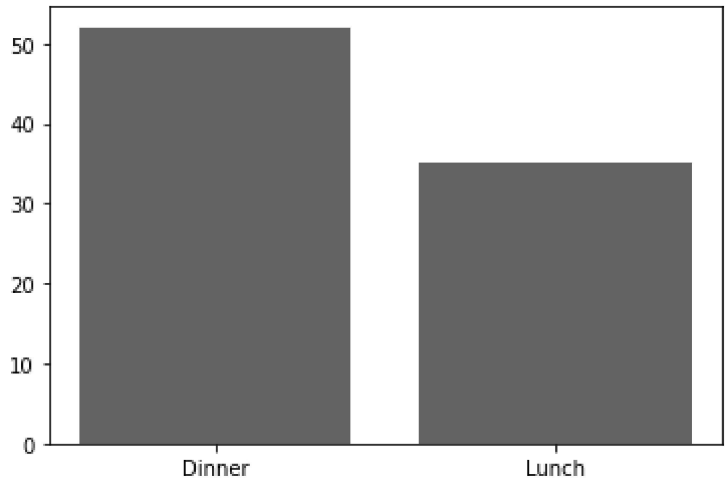
Out[48]:
```
<BarContainer object of 4 artists>
```



> The females mostly paid tips on Thursday

In [51]:
```
# Craeting a bar chart to determine which period did the females mostly paid tips
plt.bar(female_tips['time'].value_counts().index,female_tips['time'].value_counts().
```

Out[51]:
```
<BarContainer object of 2 artists>
```



> The females mostly paid tips during dinner time

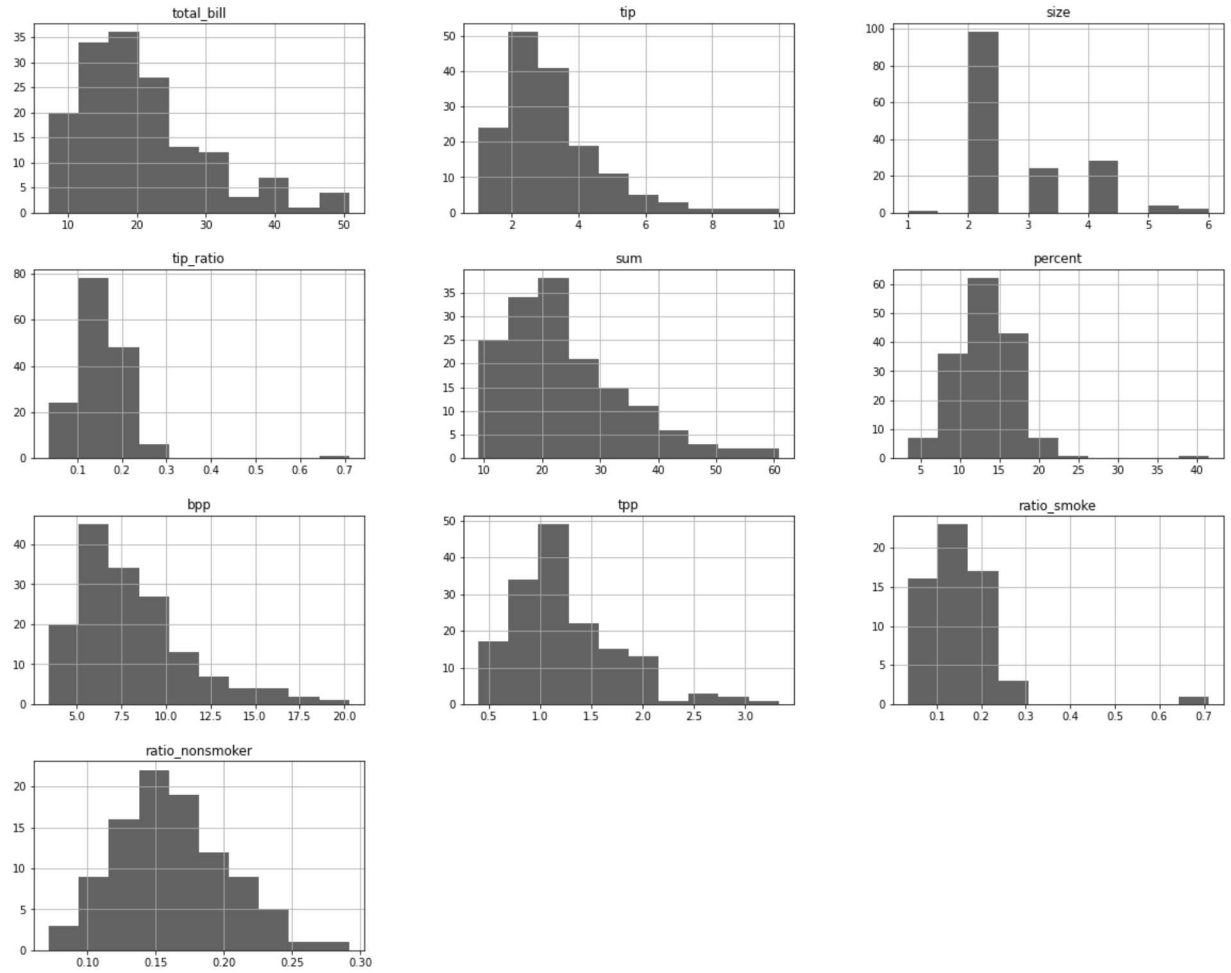
In [41]:
```
# # Subsetting the data to have only male customers
male_tips = tips.loc[tips.sex == 'Male']
male_tips.head()
```

Out[41]:

	total_bill	tip	sex	smoker	day	time	size	tip_ratio	sum	percent	bpp	tpp	r
1	10.34	1.66	Male	No	Sun	Dinner	3	0.160542	12.00	13.83	3.446667	0.553333	
2	21.01	3.50	Male	No	Sun	Dinner	3	0.166587	24.51	14.28	7.003333	1.166667	
3	23.68	3.31	Male	No	Sun	Dinner	2	0.139780	26.99	12.26	11.840000	1.655000	
5	25.29	4.71	Male	No	Sun	Dinner	4	0.186240	30.00	15.70	6.322500	1.177500	
6	8.77	2.00	Male	No	Sun	Dinner	2	0.228050	10.77	18.57	4.385000	1.000000	

In [42]:
```
# Graphical summary of the male customers
male_tips.hist(figsize=(20,16))
```

Out[42]:
```
array([[<AxesSubplot:title={'center':'total_bill'}>,
        <AxesSubplot:title={'center':'tip'}>,
        <AxesSubplot:title={'center':'size'}>],
       [<AxesSubplot:title={'center':'tip_ratio'}>,
        <AxesSubplot:title={'center':'sum'}>,
        <AxesSubplot:title={'center':'percent'}>],
       [<AxesSubplot:title={'center':'bpp'}>,
        <AxesSubplot:title={'center':'tpp'}>,
        <AxesSubplot:title={'center':'ratio_smoke'}>],
       [<AxesSubplot:title={'center':'ratio_nonsmoker'}>, <AxesSubplot:>,
        <AxesSubplot:>]], dtype=object)
```



In [43]:
```
# numerical summary of the male customers
male_tips.describe()
```

Out[43]:

	total_bill	tip	size	tip_ratio	sum	percent	bpp	tp
count	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.000000	157.00000
mean	20.744076	3.089618	2.630573	0.157651	23.833694	13.389299	8.123358	1.22282
std	9.246469	1.489102	0.955997	0.064778	10.303309	4.194238	3.119912	0.5247
min	7.250000	1.000000	1.000000	0.035638	9.000000	3.440000	3.446667	0.40000
25%	14.000000	2.000000	2.000000	0.121389	16.570000	10.820000	5.863333	0.83333
50%	18.350000	3.000000	2.000000	0.153492	21.770000	13.310000	7.600000	1.08500
75%	24.710000	3.760000	3.000000	0.186240	29.670000	15.700000	9.682500	1.50000
max	50.810000	10.000000	6.000000	0.710345	60.810000	41.530000	20.275000	3.33333

For the males:

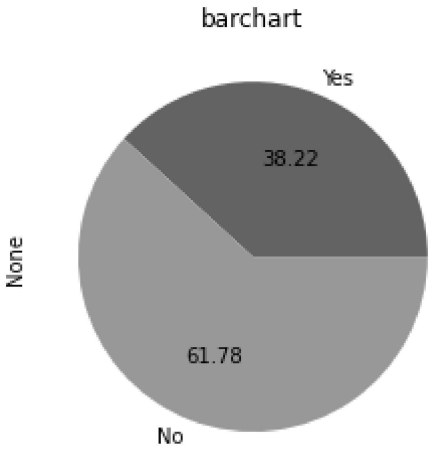
- The highest amount customers paid(excluding the tips) in the resturant was 50.81 Euros and the minimum amount was 7.25 Euros.
- The maximum amount of tips paid by customer was Euros while 10 the minimum tip paid was 1 euros.

In [46]:
```
# Getting the unique values for male smokers and non-smokers
print(male_tips.smoker.value_counts())
```

```
No     97
Yes    60
Name: smoker, dtype: int64
```

In [47]:
```
# Graphical representation of the unique values
male_tips.groupby('smoker').size().plot(kind='pie', autopct='%.2f', title = "barchar
```

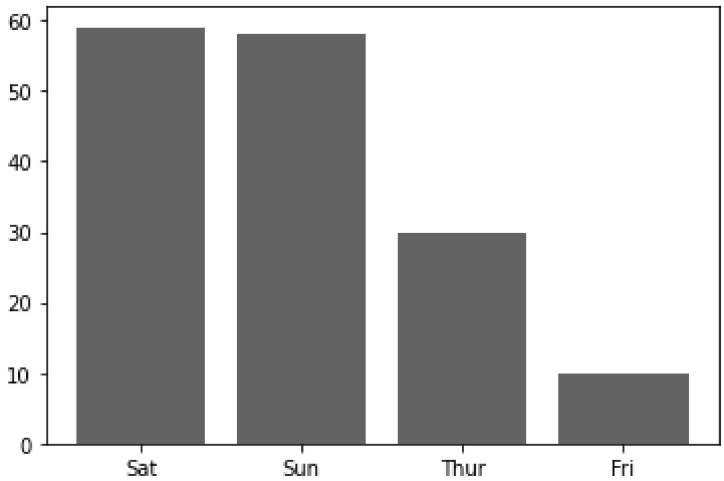
Out[47]:
```
<AxesSubplot:title={'center':'barchart'}, ylabel='None'>
```



> 61.78% of the females are non-smokers, while 38.22% are smokers

In [42]:
```
# Craeting a bar chart to determine which day did the males mostly paid tips
plt.bar(male_tips['day'].value_counts().index,male_tips['day'].value_counts().values
```

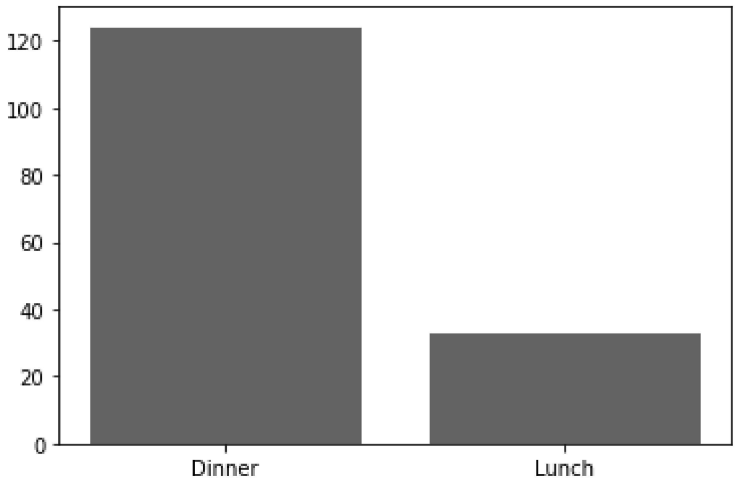
Out[42]:
```
<BarContainer object of 4 artists>
```



> The males mostly paid tips on Saturday

In [50]:
```
# Craeting a bar chart to determine which period did the males mostly paid tips
plt.bar(male_tips['time'].value_counts().index,male_tips['time'].value_counts().valu
```

Out[50]:
```
<BarContainer object of 2 artists>
```



> The males mostly paid tips during dinner time.

In [47]:
```
# Saving the data set to csv
tips.to_csv('tips.csv')
```

Summary

- On average the tips are 13.64% of the bill.
- There is a strong linear relationship between the amount of tip and the total bill for the entire dataset, the larger the bill, the larger the tip.
- As the party size increases, so does the amount of tip. However, the larger the party, the tip per person is lower.
- The males spent more money in the resturant than females.
- The smoking tippers are fewer but gives tips more.
- The females mostly paid tips on Thursday while the males mostly paid tips on Saturdays.

In [ ]:
```

```