Introduction

Welcome to **M148- Data Science Fundamentals!** This course is designed to equip you with the tools and experiences necessary to start you off on a life-long exploration of datascience. We do not assume a prerequisite knowledge or experience in order to take the course.

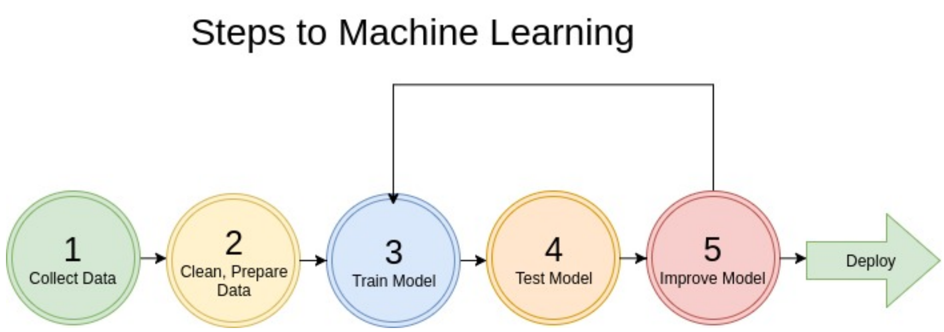
For this first project we will introduce you to the end-to-end process of doing a datascience project. Our goals for this project are to:

1. Familiarize you with the development environment for doing datascience
2. Get you comfortable with the python coding required to do datascience
3. Provide you with an sample end-to-end project to help you visualize the steps needed to complete a project on your own
4. Ask you to recreate a similar project on a separate dataset

In this project you will work through an example project end to end. Many of the concepts you will encounter will be unclear to you. That is OK! The course is designed to teach you these concepts in further detail. For now our focus is simply on having you replicate the code successfully and seeing a project through from start to finish.

Here are the main steps:

1. Get the data
2. Visualize the data for insights
3. Preprocess the data for your machine learning algorithm
4. Select a model and train
5. Does it meet the requirements? Fine tune the model



Working with Real Data

It is best to experiment with real-data as opposed to aritifical datasets.

There are many different open datasets depending on the type of problems you might be interested in!

Here are a few data repositories you could check out:

- [UCI Datasets (https://archive.ics.uci.edu/ml/)](https://archive.ics.uci.edu/ml/)
- [Kaggle Datasets (https://www.kaggle.com/)](https://www.kaggle.com/)
- [AWS Datasets (https://registry.opendata.aws)](https://registry.opendata.aws)

Submission Instructions

Project is due April 26th at 12:00 pm noon. To submit the project, please save the notebook as a pdf file and submit the assignment via Gradescope. In addition, Make sure that all figures are legible and sufficiently large.

Example Datascience Exercise

Below we will run through an California Housing example collected from the 1990's.

Setup

Before getting started, it is always good to check the versions of important packages. Knowing the version number makes it easier to lookup correct documenation.

To run this project, you will need the following packages installed with at least the minimial version number provided:

- Python Version >= 3.9
- Scitkit-learn >= 1.0.2
- Numpy >= 1.18.5
- Scipy >= 1.1.0

- Pandas >= 1.4.0
- Matplotlib >= 3.3.2

The following code imports these packages and checks their version number. If any assertion error occurs, you may not have the correct version installed.

**Important: If installed using a package manager like Anaconda or pip, these dependencies should be resolved. Please follow the python setup guide provided during discussion of week 1. **

In [1]:
```
#Import and Version Test
#Python version test
import sys
assert sys.version_info >= (3, 9) # python>=3.9

#Machine learning library
import sklearn
assert sklearn.__version__ >= "1.0.2" # sklearn >= 1.0.2

#numerical packages in python
import numpy as np
assert np.__version__ >= "1.18.5" # numpy >= 1.18.5

#Another numerical package, unused directly but is implicitly used in sklearn
#Check the version just in case
import scipy as scp
assert scp.__version__ >= "1.1.0" # scipy >= 1.1.0

#Package for data manipulation and analysis
import pandas as pd
assert pd.__version__ >= "1.4.0" # pandas >= 1.4.0

#matplotlib magic for inline figures
import matplotlib # plotting library
assert matplotlib.__version__ >= "3.3.2" # matplotlib >= 3.3.2
%matplotlib inline
```

In [2]:
```
import os
import tarfile
import urllib
DATASET_PATH = os.path.join("datasets", "housing")
```

In [3]:
```
#Other setup with necessary plotting

#Instead of using matplotlib direclty, we will use their nice pyplot interface defined as plt
import matplotlib.pyplot as plt

# Set random seed to make this notebook's output identical at every run
np.random.seed(42)

# Plotting Utilities

# Where to save the figures
ROOT_DIR = "."
IMAGES_PATH = os.path.join(ROOT_DIR, "images")
os.makedirs(IMAGES_PATH, exist_ok=True)

def save_fig(fig_name, tight_layout=True, fig_extension="png", resolution=300):
    '''
        plt.savefig wrapper. refer to
        https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.savefig.html
    '''
    path = os.path.join(IMAGES_PATH, fig_name + "." + fig_extension)
    print("Saving figure", fig_name)
    if tight_layout:
        plt.tight_layout()
    plt.savefig(path, format=fig_extension, dpi=resolution)
```

Step 1. Getting the data

Intro to Data Exploration Using Pandas

In this section we will load the dataset, do some cleaning, and visualize different features using different types of plots.

Packages we will use:

- **Pandas (https://pandas.pydata.org):** is a fast, flexibile and expressive data structure widely used for tabular and multidimensional datasets.
- **Matplotlib (https://matplotlib.org)**: is a 2d python plotting library which you can use to create quality figures (you can plot almost anything if you're willing to code it out!)

- other plotting libraries: seaborn (https://seaborn.pydata.org), ggplot2 (https://ggplot2.tidyverse.org)

In [4]:
```
import pandas as pd

def load_housing_data(housing_path):
    csv_path = os.path.join(housing_path, "housing.csv")
    return pd.read_csv(csv_path)
```

First, we load the dataset into pandas Dataframe which you can think about as an array/table. The Dataframe has a lot of useful functionality which we will use throughout the class.

In [5]:
```
housing = load_housing_data(DATASET_PATH) # we load the pandas dataframe
housing.head() # show the first few elements of the dataframe
               # typically this is the first thing you do
               # to see how the dataframe looks like
```

Out[5]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

A dataset may have different types of features

- real valued
- Discrete (integers)
- categorical (strings)

The two categorical features are essentialy the same as you can always map a categorical string/character to an integer.

In the dataset example, all our features are real valued floats, except ocean proximity which is categorical.

In [6]:
```
# to see a concise summary of data types, null values, and counts
# use the info() method on the dataframe
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20640 non-null  float64
 3   total_rooms         20640 non-null  float64
 4   total_bedrooms      20433 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  float64
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  float64
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

In [7]:
```
# you can access individual columns similarly
# to accessing elements in a python dict
print(housing["ocean_proximity"].head()) # added head() to avoid printing many columns.

#Additionally, columns can be accessed as attirbutes of the dataframe object
#This method is convenient to access data but should be used with care since you can't overwrite
#built in functions like housing.min()
print(housing.ocean_proximity.head())
```

```
0    NEAR BAY
1    NEAR BAY
2    NEAR BAY
3    NEAR BAY
4    NEAR BAY
Name: ocean_proximity, dtype: object
0    NEAR BAY
1    NEAR BAY
2    NEAR BAY
3    NEAR BAY
4    NEAR BAY
Name: ocean_proximity, dtype: object
```

In [8]:
```
# to access a particular row we can use iloc
housing.iloc[1]
```

Out[8]:
```
longitude              -122.22
latitude                 37.86
housing_median_age        21.0
total_rooms             7099.0
total_bedrooms          1106.0
population              2401.0
households              1138.0
median_income           8.3014
median_house_value    358500.0
ocean_proximity       NEAR BAY
Name: 1, dtype: object
```

In [9]:
```
# one other function that might be useful is
# value_counts(), which counts the number of occurences
# for categorical features
housing["ocean_proximity"].value_counts()
```

Out[9]:
```
<1H OCEAN     9136
INLAND        6551
NEAR OCEAN    2658
NEAR BAY      2290
ISLAND           5
Name: ocean_proximity, dtype: int64
```

In [10]:
```
# The describe function compiles your typical statistics for each non-categorical column
housing.describe()
```

Out[10]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

We can also perform groupings based on categorical values and analyze each group.

In [11]:
```
housing_group = housing.groupby('ocean_proximity')
#Has the mean for every column grouped by ocean proximity
housing_mean = housing_group.mean()
housing_mean
```

Out[11]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
ocean_proximity									
<1H OCEAN	-118.847766	34.560577	29.279225	2628.343586	546.539185	1520.290499	517.744965	4.230682	240084.285464
INLAND	-119.732990	36.731829	24.271867	2717.742787	533.881619	1391.046252	477.447565	3.208996	124805.392001
ISLAND	-118.354000	33.358000	42.400000	1574.600000	420.400000	668.000000	276.600000	2.744420	380440.000000
NEAR BAY	-122.260694	37.801057	37.730131	2493.589520	514.182819	1230.317467	488.616157	4.172885	259212.311790
NEAR OCEAN	-119.332555	34.738439	29.347254	2583.700903	538.615677	1354.008653	501.244545	4.005785	249433.977427

In [12]:
```
#We can also get the subset of data associated with that group

housing_inland = housing_group.get_group("INLAND")
housing_inland
```

Out[12]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
954	-121.92	37.64	46.0	1280.0	209.0	512.0	208.0	5.1406	315600.0	INLAND
957	-121.90	37.66	18.0	7397.0	1137.0	3126.0	1115.0	6.4994	323000.0	INLAND
965	-121.88	37.68	23.0	2234.0	270.0	854.0	286.0	7.3330	337200.0	INLAND
967	-121.88	37.67	16.0	4070.0	624.0	1543.0	577.0	6.5214	311500.0	INLAND
968	-121.88	37.67	25.0	2244.0	301.0	937.0	324.0	6.4524	296900.0	INLAND
...	...	...	...	...	...	...	...	...	...	...
20635	-121.09	39.48	25.0	1665.0	374.0	845.0	330.0	1.5603	78100.0	INLAND
20636	-121.21	39.49	18.0	697.0	150.0	356.0	114.0	2.5568	77100.0	INLAND
20637	-121.22	39.43	17.0	2254.0	485.0	1007.0	433.0	1.7000	92300.0	INLAND
20638	-121.32	39.43	18.0	1860.0	409.0	741.0	349.0	1.8672	84700.0	INLAND
20639	-121.24	39.37	16.0	2785.0	616.0	1387.0	530.0	2.3886	89400.0	INLAND

6551 rows × 10 columns

In [13]:
```
#We can thus performs operations on each group separately
housing_inland.describe()
```

Out[13]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
count	6551.00000	6551.000000	6551.000000	6551.000000	6496.000000	6551.000000	6551.000000	6551.000000	6551.000000
mean	-119.73299	36.731829	24.271867	2717.742787	533.881619	1391.046252	477.447565	3.208996	124805.392001
std	1.90095	2.116073	12.018020	2385.831111	446.117778	1168.670126	392.252095	1.437465	70007.908494
min	-123.73000	32.640000	1.000000	2.000000	2.000000	5.000000	2.000000	0.499900	14999.000000
25%	-121.35000	34.180000	15.000000	1404.000000	282.000000	722.000000	254.000000	2.188950	77500.000000
50%	-120.00000	36.970000	23.000000	2131.000000	423.000000	1124.000000	385.000000	2.987700	108500.000000
75%	-117.84000	38.550000	33.000000	3216.000000	636.000000	1687.000000	578.000000	3.961500	148950.000000
max	-114.31000	41.950000	52.000000	39320.000000	6210.000000	16305.000000	5358.000000	15.000100	500001.000000

Grouping is a powerful technique within pandas and a recommend reading the user guide to understand it better [here (https://pandas.pydata.org/docs/user_guide/groupby.html)](https://pandas.pydata.org/docs/user_guide/groupby.html)

In addition to grouping, we can also filter out the data based on our desired criteria.

In [14]:
```
housing_expensive= housing[(housing["median_house_value"] > 50000)]
housing_expensive.head()
```

Out[14]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

In [15]:
```
#We can combine multiple criteria
housing_expensive_small= housing[(housing["median_house_value"] > 50000)& (housing["population"] < 1000)]
housing_expensive_small.head()
```

Out[15]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	NEAR BAY

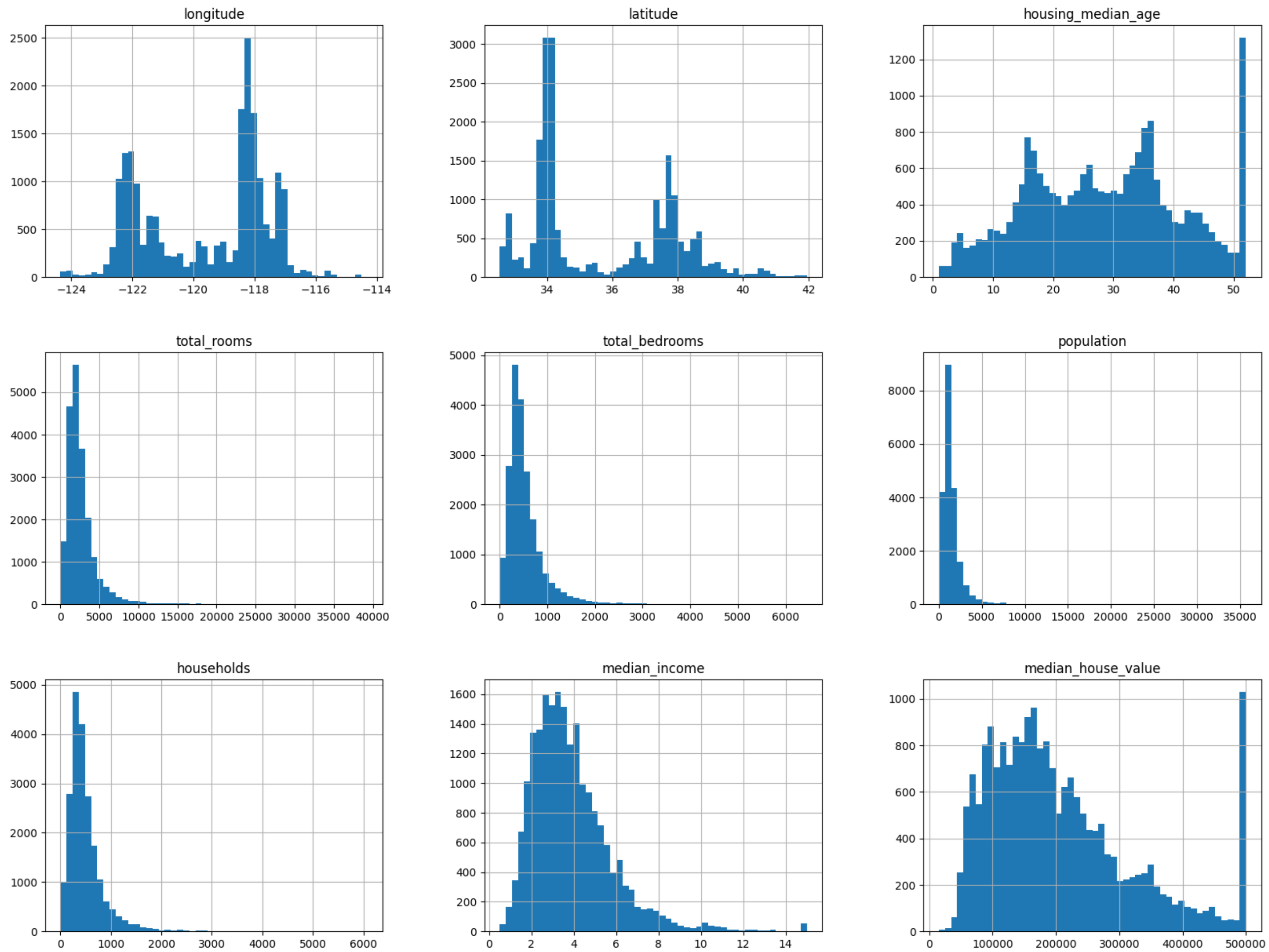
If you want to learn about different ways of accessing elements or other functions it's useful to check out the getting started section of pandas [here (https://pandas.pydata.org/pandas-docs/stable/getting_started/index.html)](https://pandas.pydata.org/pandas-docs/stable/getting_started/index.html) and for a full look at all the functionaltiy that pandas offers you can check out the user guide of pandas [here (https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html)](https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html)

Step 2. Visualizing the data

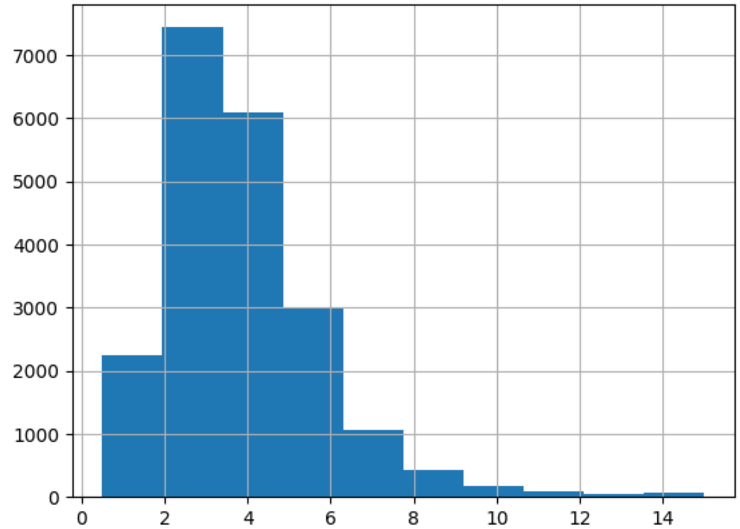
Let's start visualizing the dataset

In [16]:

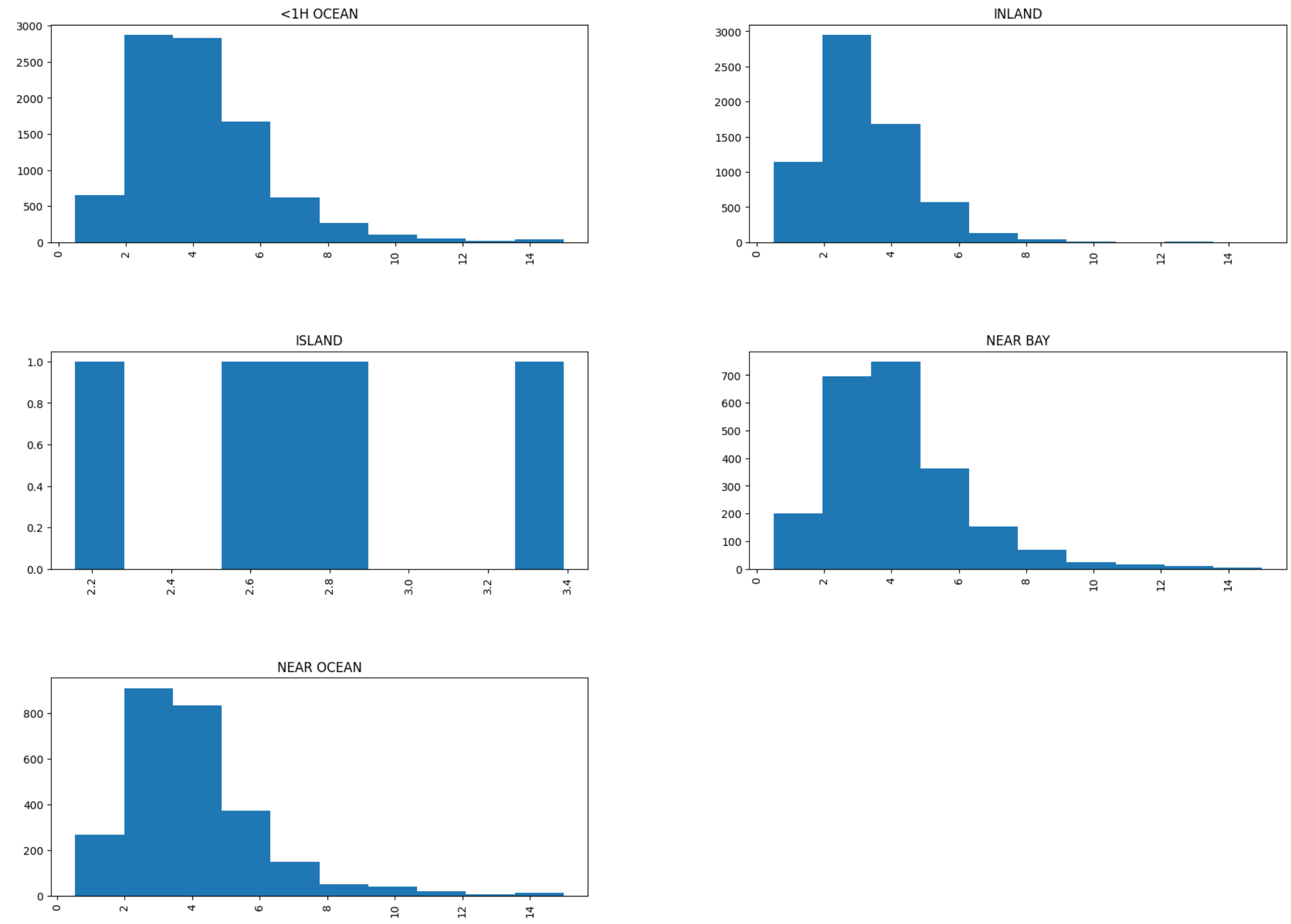
```
# We can draw a histogram for each of the dataframes features
# using the built-in hist function of Dataframe
housing.hist(bins=50, figsize=(20,15))
#save_fig("attribute_histogram_plots")
plt.show() # pandas internally uses matplotlib, and to display all the figures
           # the show() function must be called
```



In [17]:
```
# if you want to have a histogram on an individual feature:
housing["median_income"].hist()
plt.show()
```

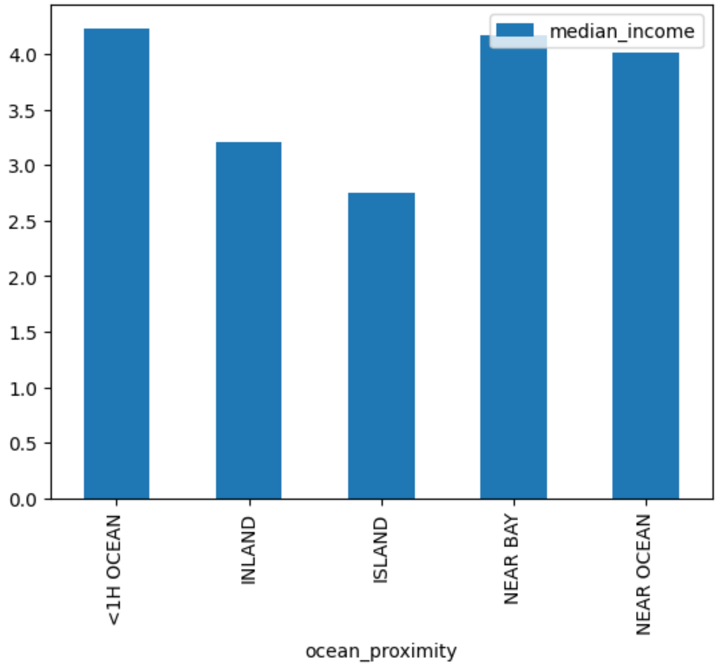


In [18]:
```
#You can even plot histograms by specifying the groupings using by
housing["median_income"].hist(by= housing["ocean_proximity"],figsize=(20,15))
plt.show()
```



In [19]:
```
#We can also plot statistics of each groupings
housing_group_mean = housing.groupby("ocean_proximity").mean()

housing_group_mean.plot.bar(y ="median_income")
```

Out[19]:
```
<AxesSubplot: xlabel='ocean_proximity'>
```



We can convert a floating point feature to a categorical feature by binning or by defining a set of intervals.

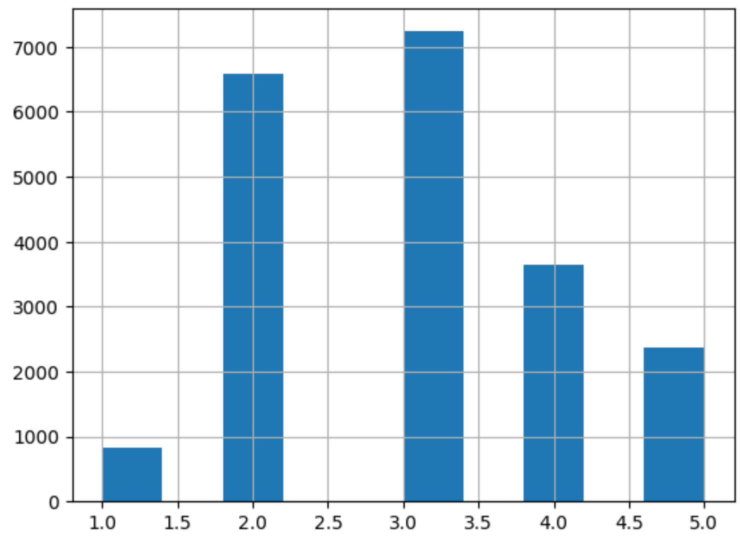
For example, to bin the households based on median_income we can use the pd.cut function. Note that we use np.inf to represent infinity which is internally handeled. Thus, the last bin is $(6, \infty)$.

In [20]:
```
# assign each bin a categorical value [1, 2, 3, 4, 5] in this case.
housing["income_cat"] = pd.cut(housing["median_income"],
                               bins=[0., 1.5, 3.0, 4.5, 6., np.inf],
                               labels=[1, 2, 3, 4, 5])

housing["income_cat"].value_counts()
```

Out[20]:
```
3    7236
2    6581
4    3639
5    2362
1     822
Name: income_cat, dtype: int64
```

In [21]:
```
housing["income_cat"].hist()
```

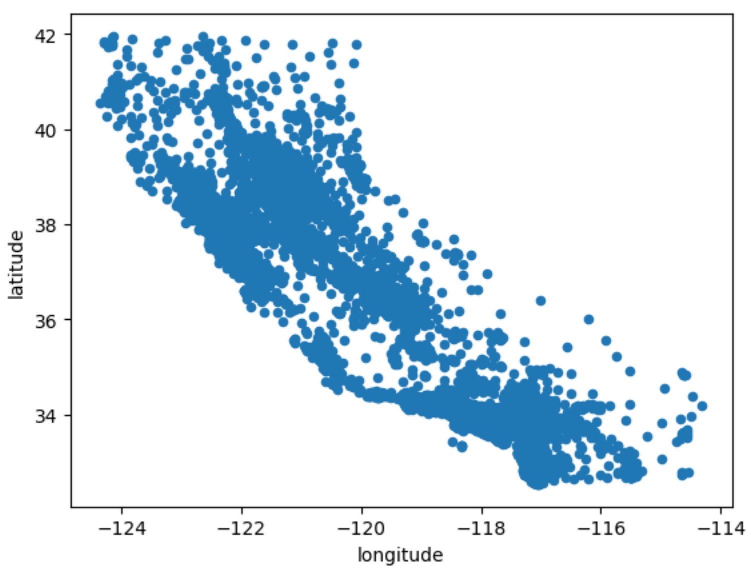
Out[21]: <AxesSubplot: >



Next let's visualize the household incomes based on latitude & longitude coordinates

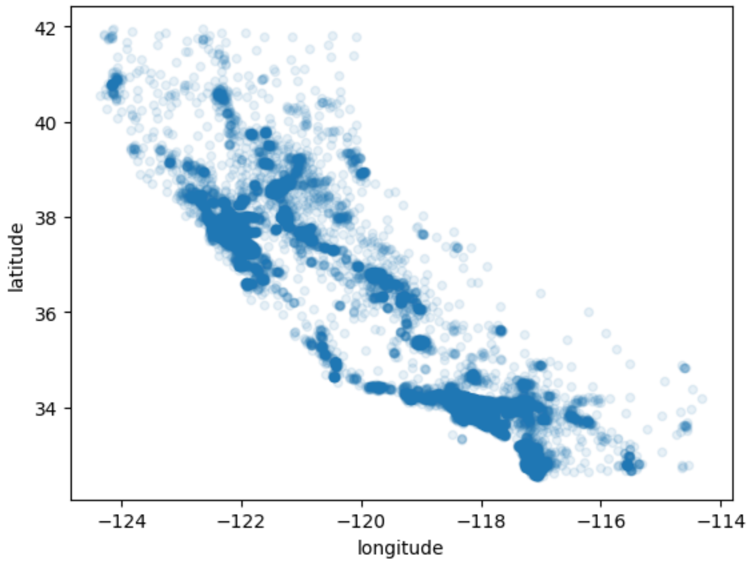
In [22]:
```
## here's a not so interesting way of plotting it
housing.plot(kind="scatter", x="longitude", y="latitude")
#save_fig("bad_visualization_plot")
```

Out[22]: <AxesSubplot: xlabel='longitude', ylabel='latitude'>



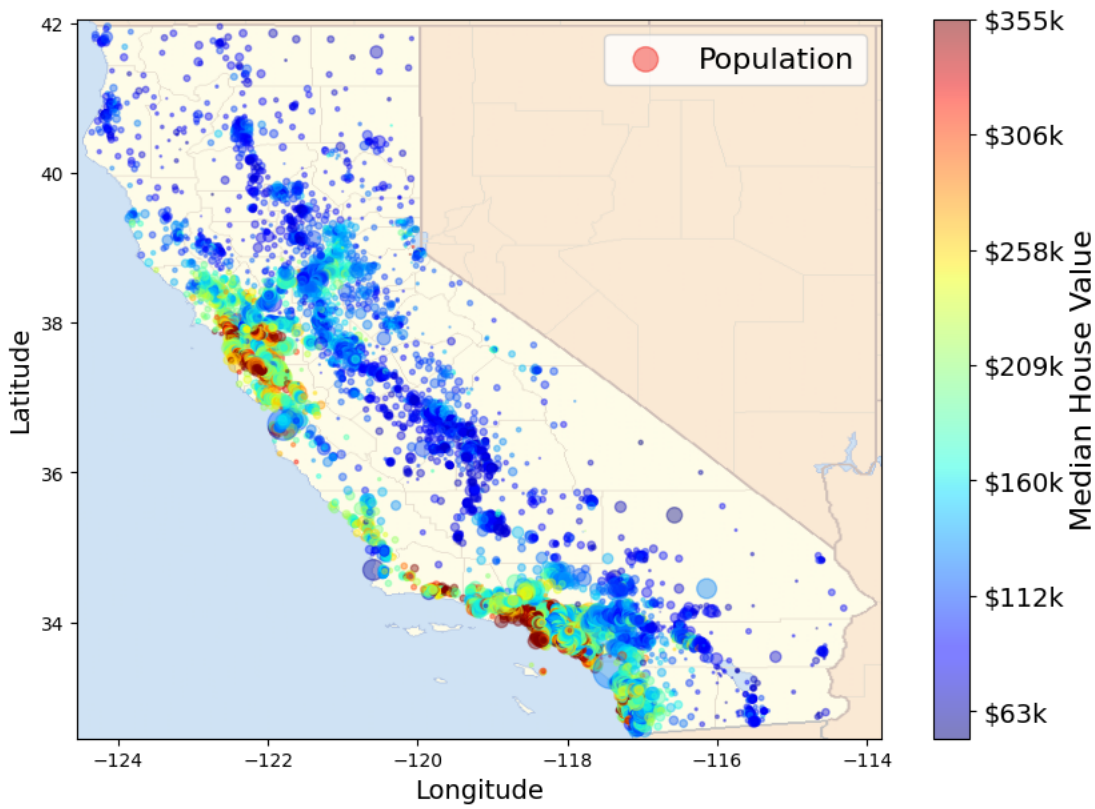
In [23]:
```
# we can make it look a bit nicer by using the alpha parameter,
# it simply plots less dense areas lighter.
housing.plot(kind="scatter", x="longitude", y="latitude", alpha=0.1)
#save_fig("better_visualization_plot")
```

Out[23]: <AxesSubplot: xlabel='longitude', ylabel='latitude'>



In [24]:
```
# A more interesting plot is to color code (heatmap) the dots
# based on income. The code below achieves this

# load an image of california
images_path = os.path.join('./', "images")
os.makedirs(images_path, exist_ok=True)
filename = "california.png"

import matplotlib.image as mpimg
california_img=mpimg.imread(os.path.join(images_path, filename))
"""ax = housing.plot(kind="scatter", x="longitude", y="latitude", figsize=(10,7),
                       s=housing['population']/100, label="Population",
                       c="median_house_value", cmap=plt.get_cmap("jet"),
                       colorbar=False, alpha=0.4,
                      )
"""
ax = housing.plot(kind="scatter", x="longitude", y="latitude", figsize=(10,7),
                       s=housing['population']/100, label="Population",
                       c="median_house_value", cmap="jet",
                       colorbar=False, alpha=0.4,
                      )
# overlay the califronia map on the plotted scatter plot
# note: plt.imshow still refers to the most recent figure
# that hasn't been plotted yet.
plt.imshow(california_img, extent=[-124.55, -113.80, 32.45, 42.05], alpha=0.5,
           cmap=plt.get_cmap("jet"))
plt.ylabel("Latitude", fontsize=14)
plt.xlabel("Longitude", fontsize=14)

# setting up heatmap colors based on median_house_value feature
prices = housing["median_house_value"]
tick_values = np.linspace(prices.min(), prices.max(), 11)
cb = plt.colorbar()
cb.ax.set_yticklabels(["$%dk"%(round(v/1000)) for v in tick_values], fontsize=14)
cb.set_label('Median House Value', fontsize=16)

plt.legend(fontsize=16)
#save_fig("california_housing_prices_plot")
plt.show()
```

```
/tmp/ipykernel_10764/715432031.py:34: UserWarning: FixedFormatter should only be used together with FixedLocator
  cb.ax.set_yticklabels(["$%dk"%(round(v/1000)) for v in tick_values], fontsize=14)
```



Not suprisingly, we can see that the most expensive houses are concentrated around the San Francisco/Los Angeles areas.

Up until now we have only visualized feature histograms and basic statistics.

When developing machine learning models the predictiveness of a feature for a particular target of interest is what's important.

It may be that only a few features are useful for the target at hand, or features may need to be augmented by applying certain transformations.

Nonetheless we can explore this using correlation matrices. Each row and column of the correlation matrix represents a non-categorical feature in our dataset and each element specifies the correlation between the row and column features. [Correlation (https://en.wikipedia.org/wiki/Correlation)](https://en.wikipedia.org/wiki/Correlation) is a measure of how the change in one feature affects the other feature. For example, a positive correlation means that as one feature gets larger, then the other feature will also generally get larger. Note that a feature is always fully correlated to itself which is why the diagonal of the correlation matrix is just all 1s.

In [25]:
```
corr_matrix = housing.corr()
corr_matrix
```

```
/tmp/ipykernel_10764/1253314489.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprec
ated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_onl
y to silence this warning.
  corr_matrix = housing.corr()
```

Out[25]:

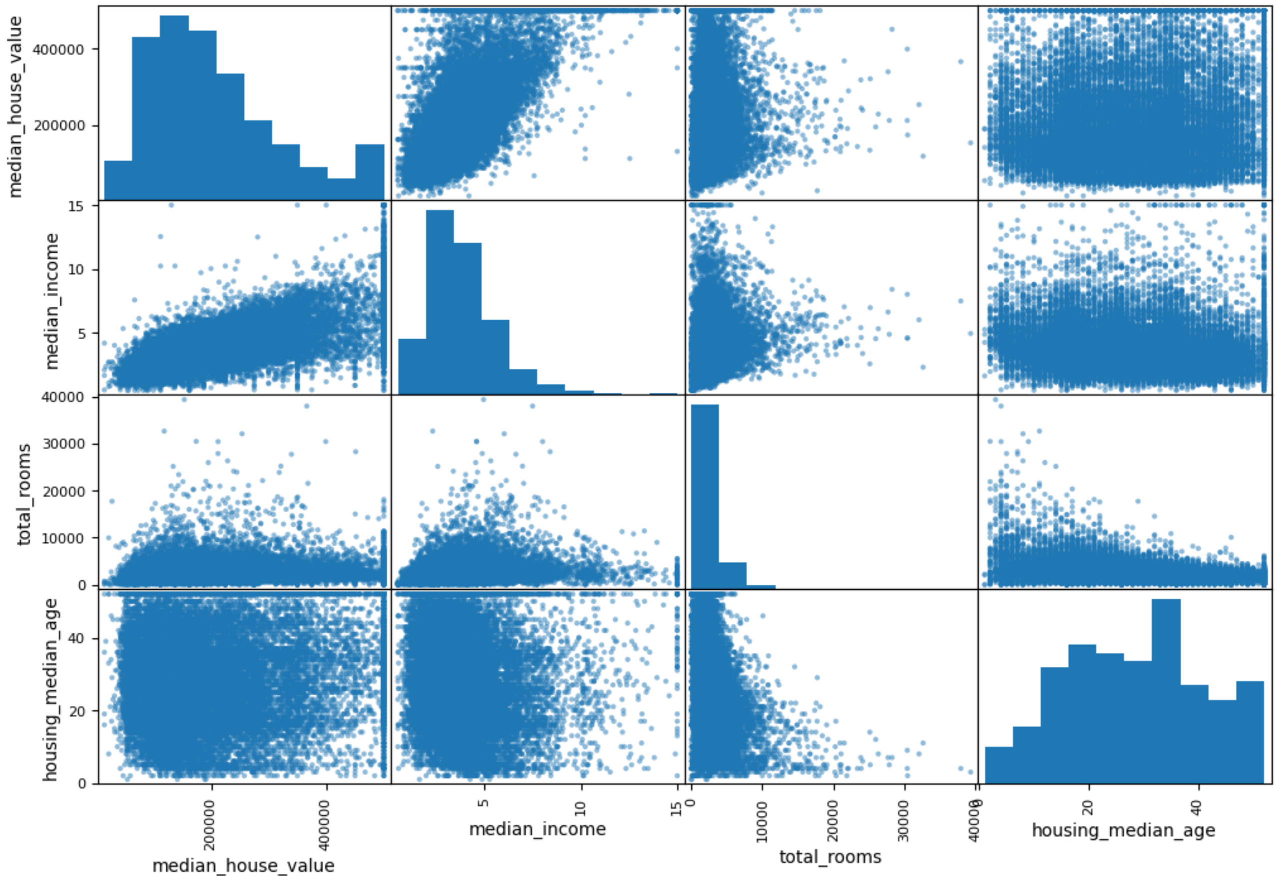
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
longitude	1.000000	-0.924664	-0.108197	0.044568	0.069608	0.099773	0.055310	-0.015176	-0.045967
latitude	-0.924664	1.000000	0.011173	-0.036100	-0.066983	-0.108785	-0.071035	-0.079809	-0.144160
housing_median_age	-0.108197	0.011173	1.000000	-0.361262	-0.320451	-0.296244	-0.302916	-0.119034	0.105623
total_rooms	0.044568	-0.036100	-0.361262	1.000000	0.930380	0.857126	0.918484	0.198050	0.134153
total_bedrooms	0.069608	-0.066983	-0.320451	0.930380	1.000000	0.877747	0.979728	-0.007723	0.049686
population	0.099773	-0.108785	-0.296244	0.857126	0.877747	1.000000	0.907222	0.004834	-0.024650
households	0.055310	-0.071035	-0.302916	0.918484	0.979728	0.907222	1.000000	0.013033	0.065843
median_income	-0.015176	-0.079809	-0.119034	0.198050	-0.007723	0.004834	0.013033	1.000000	0.688075
median_house_value	-0.045967	-0.144160	0.105623	0.134153	0.049686	-0.024650	0.065843	0.688075	1.000000

In [26]:
```
# for example if the target is "median_house_value", most correlated features can be sorted
# which happens to be "median_income". This also intuitively makes sense.
corr_matrix["median_house_value"].sort_values(ascending=False)
```

Out[26]:
```
median_house_value    1.000000
median_income         0.688075
total_rooms           0.134153
housing_median_age    0.105623
households            0.065843
total_bedrooms        0.049686
population           -0.024650
longitude            -0.045967
latitude             -0.144160
Name: median_house_value, dtype: float64
```

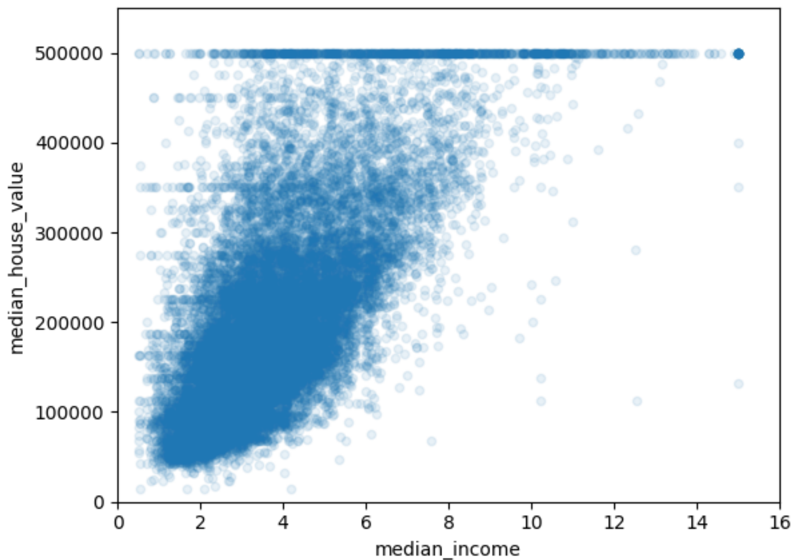
In [27]:
```
# We can plot a scatter matrix for different attributes/features
# to see how some features may show a positive correlation/negative correlation or
# it may turn out to be completely random!
from pandas.plotting import scatter_matrix
attributes = ["median_house_value", "median_income", "total_rooms",
              "housing_median_age"]
scatter_matrix(housing[attributes], figsize=(12, 8))
#save_fig("scatter_matrix_plot")
```

Out[27]:
```
array([[<AxesSubplot: xlabel='median_house_value', ylabel='median_house_value'>,
        <AxesSubplot: xlabel='median_income', ylabel='median_house_value'>,
        <AxesSubplot: xlabel='total_rooms', ylabel='median_house_value'>,
        <AxesSubplot: xlabel='housing_median_age', ylabel='median_house_value'>],
       [<AxesSubplot: xlabel='median_house_value', ylabel='median_income'>,
        <AxesSubplot: xlabel='median_income', ylabel='median_income'>,
        <AxesSubplot: xlabel='total_rooms', ylabel='median_income'>,
        <AxesSubplot: xlabel='housing_median_age', ylabel='median_income'>],
       [<AxesSubplot: xlabel='median_house_value', ylabel='total_rooms'>,
        <AxesSubplot: xlabel='median_income', ylabel='total_rooms'>,
        <AxesSubplot: xlabel='total_rooms', ylabel='total_rooms'>,
        <AxesSubplot: xlabel='housing_median_age', ylabel='total_rooms'>],
       [<AxesSubplot: xlabel='median_house_value', ylabel='housing_median_age'>,
        <AxesSubplot: xlabel='median_income', ylabel='housing_median_age'>,
        <AxesSubplot: xlabel='total_rooms', ylabel='housing_median_age'>,
        <AxesSubplot: xlabel='housing_median_age', ylabel='housing_median_age'>]],
      dtype=object)
```



In [28]:
```
# median income vs median house value plot 2 in the first row of top figure
housing.plot(kind="scatter", x="median_income", y="median_house_value",
             alpha=0.1)
plt.axis([0, 16, 0, 550000])
#save_fig("income_vs_house_value_scatterplot")
```

Out[28]:
```
(0.0, 16.0, 0.0, 550000.0)
```



Augmenting Features: Simple Example

New features can be created by combining different columns from our data set.

- rooms_per_household = total_rooms / households
- bedrooms_per_room = total_bedrooms / total_rooms
- etc.

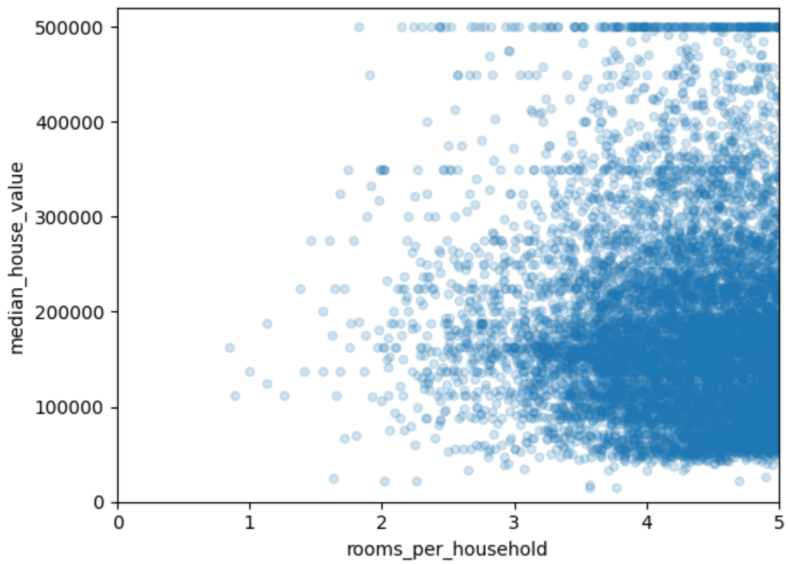
In [29]:
```
#A new column in the dataframe can be made the same away you add a new element to a dict
housing["rooms_per_household"] = housing["total_rooms"]/housing["households"]
housing["bedrooms_per_room"] = housing["total_bedrooms"]/housing["total_rooms"]
housing["population_per_household"]=housing["population"]/housing["households"]
```

In [30]:
```
# obtain new correlations
corr_matrix = housing.corr()
corr_matrix["median_house_value"].sort_values(ascending=False)
```

```
/tmp/ipykernel_10764/313240856.py:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  corr_matrix = housing.corr()
```

Out[30]:
```
median_house_value          1.000000
median_income               0.688075
rooms_per_household         0.151948
total_rooms                 0.134153
housing_median_age          0.105623
households                  0.065843
total_bedrooms              0.049686
population_per_household   -0.023737
population                 -0.024650
longitude                  -0.045967
latitude                   -0.144160
bedrooms_per_room          -0.255880
Name: median_house_value, dtype: float64
```

In [31]:
```
housing.plot(kind="scatter", x="rooms_per_household", y="median_house_value",
             alpha=0.2)
plt.axis([0, 5, 0, 520000])
plt.show()
```



In [32]:
```
housing.describe()
```

Out[32]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	r
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000	
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909	
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874	
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000	
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000	
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000	
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000	
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000	

Augmenting Features: Advanced Example

In addition to augmenting the data using these simple operations, we can also do some advanced augmentation by bringing information from another dataset.

In this case, we are going to find the distance between the houses and the 10 biggest cities in California during 1990. Intuitively, the location of major cities can strongly impact the value of a home. Thus, our new feature will be the distance of the home to the closest big city among the 10 biggest cities.

To perform this feature extraction, we will use the provided dataset "city_data.csv". We will also employ some helper functions and use the pd.apply function to do the augmentation.

In [33]:
```
#Loads the city data
def load_city_data(housing_path):
    csv_path = os.path.join(housing_path, "city_data.csv")
    return pd.read_csv(csv_path)

city_data = load_city_data(DATASET_PATH)
city_data
```

Out[33]:

	City	Latitude	Longitude	Pop_1990
0	Anaheim	33.835292	-117.914503	266406
1	Fresno	36.746842	-119.772586	354202
2	Long Beach	33.768322	-118.195617	429433
3	Los Angeles	34.052233	-118.243686	3485398
4	Oakland	37.804364	-122.271114	372242
5	Sacramento	38.581572	-121.494400	369365
6	San Diego	32.715328	-117.157256	1110549
7	San Francisco	37.774931	-122.419417	723959
8	San Jose	37.339386	-121.894956	782248
9	Santa Ana	33.745572	-117.867833	293742

In [34]:
```
#For ease of use, we will convert city_data into a python dict
#where the key is the city name and the value is the coordinates
city_dict = {}
for dat in city_data.iterrows(): #iterates through the rows of the dataframe
    row = dat[1]
    city_dict[row["City"]] = (row["Latitude"],row["Longitude"])

print(city_dict)
```

```
{'Anaheim': (33.835292, -117.914503), 'Fresno': (36.746842, -119.772586), 'Long Beach': (33.768322, -118.195617),
'Los Angeles': (34.052233, -118.243686), 'Oakland': (37.804364, -122.271114), 'Sacramento': (38.581572, -121.494
4), 'San Diego': (32.715328, -117.157256), 'San Francisco': (37.774931, -122.419417), 'San Jose': (37.339386, -12
1.894956), 'Santa Ana': (33.745572, -117.867833)}
```

In [35]:
```
#Helper functions

#This function is used to calculate the distance between two points on a latitude and longitude grid.
#You don't need to understand the math but know that it takes into account the curverature of the earth
#to make an accurate distance measurement.
#While we could have used the geopy package to do this for us, this way we don't have to install it.
def distance_func(loc_a,loc_b):
    """
    Calculates the haversine distance between coordinates
    on the latitude and longitude grid.
    Distance is in km.
    """
    lat1,lon1 = loc_a
    lat2,lon2 = loc_b
    r = 6371
    phi1 = np.radians(lat1)
    phi2 = np.radians(lat2)
    delta_phi = np.radians(lat2 - lat1)
    delta_lambda = np.radians(lon2 - lon1)
    a = np.sin(delta_phi / 2)**2 + np.cos(phi1) * np.cos(phi2) *   np.sin(delta_lambda / 2)**2
    res = r * (2 * np.arctan2(np.sqrt(a), np.sqrt(1 - a)))
    return np.round(res, 2)


#Calculates closest point to the location given in kilometers
def closest_point(location, location_dict):
    """ take a tuple of latitude and longitude and
        compare to a dictionary of locations where
        key = location name and value = (lat, long)
        returns tuple of (closest_location , distance)
        distance is in kilometers"""
    closest_location = None
    for city in location_dict.keys():
        distance = distance_func(location, location_dict[city])
        if closest_location is None:
            closest_location = (city, distance)
        elif distance < closest_location[1]:
            closest_location = (city, distance)
    return closest_location

#Example
closest_point((37.774931,-120.419417), city_dict)
```

Out[35]: ('Fresno', 127.85)

In [36]:
```
#Now we apply the closest_point function to every data point in housing
#Axis = 1 specifies that apply will send each row one by one into the designated function
#We use the lambda function to catch the row and then disperse its arguments into closest_point
housing['close_city'] = housing.apply(lambda x: closest_point((x['latitude'],x['longitude']),city_dict), axis = 1)

#Since closest point outputted a tuple of names and distance, we have to split it up.
housing['close_city_name'] = [x[0] for x in housing['close_city'].values]
housing['close_city_dist'] = [x[1] for x in housing['close_city'].values]

#Drop the redundant column
housing = housing.drop('close_city', axis=1)
```

In [37]:
```
#Now, let us look at our new features
housing.head()
```

Out[37]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity	in
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY	
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY	
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY	
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY	
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY	

In [38]:
```
#We can also look at the new statistics
housing.describe()
```

Out[38]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	r
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000	
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909	
std	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874	
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000	
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000	
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000	
75%	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000	
max	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000	

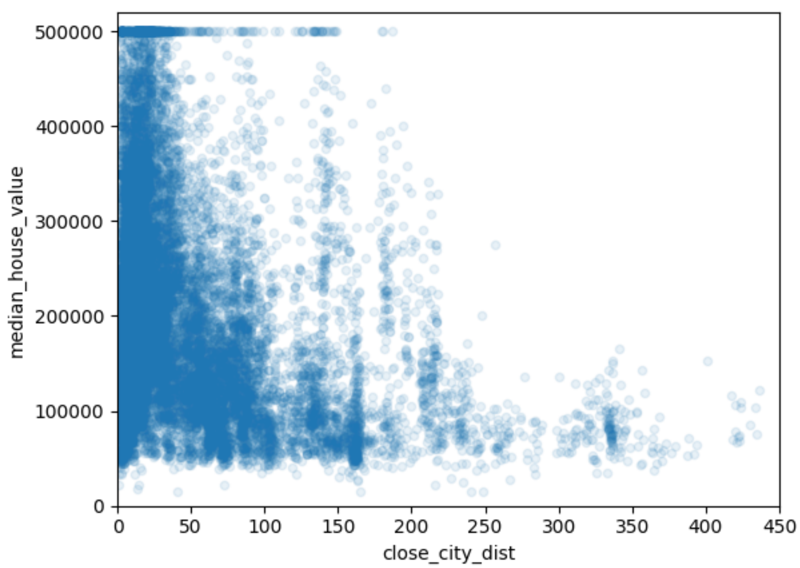
Now, let us see if the new feature provides some information about housing prices by looking at the correlation.

In [39]:
```
# obtain new correlations
corr_matrix = housing.corr()
corr_matrix["median_house_value"].sort_values(ascending=False)
```

```
/tmp/ipykernel_10764/313240856.py:2: FutureWarning: The default value of numeric_only in DataFrame.corr is depreca
ted. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  corr_matrix = housing.corr()
```

Out[39]:
```
median_house_value          1.000000
median_income               0.688075
rooms_per_household         0.151948
total_rooms                 0.134153
housing_median_age          0.105623
households                  0.065843
total_bedrooms              0.049686
population_per_household   -0.023737
population                 -0.024650
longitude                  -0.045967
latitude                   -0.144160
bedrooms_per_room          -0.255880
close_city_dist            -0.307777
Name: median_house_value, dtype: float64
```

In [40]:
```
housing.plot(kind="scatter", x="close_city_dist", y="median_house_value",
             alpha=0.1)
plt.axis([0, 450, 0, 520000])
plt.show()
```



Observation: From the correlation, we can see a negative correlation implying that the farther a house is from a big city, the less it costs. From the plot, we can confirm the negative correlation. We can also note that most houses are within 250 km of the big cities which can indicate that everything past 250 is an outlier or should be treated differently like farm land.

Step 3. Preprocess the data for your machine learning algorithm

Once we've visualized the data, and have a certain understanding of how the data looks like. It's time to clean!

Most of your time will be spent on this step, although the datasets used in this project are relatively nice and clean... in the real world it could get real dirty.

After having cleaned your dataset you're aiming for:

- train set
- test set

In some cases you might also have a validation set as well for tuning hyperparameters (don't worry if you're not familiar with this term yet..)

In supervised learning setting your train set and test set should contain (**feature**, **target**) tuples.

- **feature**: is the input to your model
- **target**: is the ground truth label
  - when target is categorical the task is a classification task
  - when target is floating point the task is a regression task

We will make use of **scikit-learn (https://scikit-learn.org/stable/)** python package for preprocessing.

Scikit learn is pretty well documented and if you get confused at any point simply look up the function/object here (https://scikit-learn.org/stable/user_guide.html)!

Dealing With Incomplete Data

In [41]:
```
# have you noticed when looking at the dataframe summary certain rows
# contained null values? we can't just leave them as nulls and expect our
# model to handle them for us so we'll have to devise a method for dealing with them...
sample_incomplete_rows = housing[housing.isnull().any(axis=1)].head()
sample_incomplete_rows
```

Out[41]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
290	-122.16	37.77	47.0	1256.0	NaN	570.0	218.0	4.3750	161900.0	NEAR BAY
341	-122.17	37.75	38.0	992.0	NaN	732.0	259.0	1.6196	85100.0	NEAR BAY
538	-122.28	37.78	29.0	5154.0	NaN	3741.0	1273.0	2.5762	173400.0	NEAR BAY
563	-122.24	37.75	45.0	891.0	NaN	384.0	146.0	4.9489	247100.0	NEAR BAY
696	-122.10	37.69	41.0	746.0	NaN	387.0	161.0	3.9063	178400.0	NEAR BAY

In [42]:
```
sample_incomplete_rows.dropna(subset=["total_bedrooms"])    # option 1: simply drop rows that have null values
```

Out[42]:

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity	inco

In [43]:
```
sample_incomplete_rows.drop("total_bedrooms", axis=1)       # option 2: drop the complete feature
```

Out[43]:

	longitude	latitude	housing_median_age	total_rooms	population	households	median_income	median_house_value	ocean_proximity	income_cat	roon
290	-122.16	37.77	47.0	1256.0	570.0	218.0	4.3750	161900.0	NEAR BAY	3	
341	-122.17	37.75	38.0	992.0	732.0	259.0	1.6196	85100.0	NEAR BAY	2	
538	-122.28	37.78	29.0	5154.0	3741.0	1273.0	2.5762	173400.0	NEAR BAY	2	
563	-122.24	37.75	45.0	891.0	384.0	146.0	4.9489	247100.0	NEAR BAY	4	
696	-122.10	37.69	41.0	746.0	387.0	161.0	3.9063	178400.0	NEAR BAY	3	

In [44]:
```
median = housing["total_bedrooms"].median()
sample_incomplete_rows["total_bedrooms"].fillna(median, inplace=True) # option 3: replace na values with median valu
sample_incomplete_rows
```

Out[44]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
290	-122.16	37.77	47.0	1256.0	435.0	570.0	218.0	4.3750	161900.0	NEAR BAY
341	-122.17	37.75	38.0	992.0	435.0	732.0	259.0	1.6196	85100.0	NEAR BAY
538	-122.28	37.78	29.0	5154.0	435.0	3741.0	1273.0	2.5762	173400.0	NEAR BAY
563	-122.24	37.75	45.0	891.0	435.0	384.0	146.0	4.9489	247100.0	NEAR BAY
696	-122.10	37.69	41.0	746.0	435.0	387.0	161.0	3.9063	178400.0	NEAR BAY

The option where we replace the null values with a new number is known as [imputation (https://en.wikipedia.org/wiki/Imputation_(statistics))](https://en.wikipedia.org/wiki/Imputation_(statistics)).

Could you think of another plausible imputation for this dataset instead of using the median? (Not graded)

Using Scikit-learn transformers to preprocess data

We have shown some operations that we want to perform on the dataset. While it is possible to manually perform it all yourselves, it is much easier to offload some of the work to the many fantastic machine learning packages. One such example is scikit-learn where we will demonstrate the use of a transformer to handle some of the work.

Consider a situation where we want to normalize the data for each feature. This involves calculating the mean μ and standard deviation σ for that feature and applying $\frac{z-\mu}{\sigma}$ where z is the feature value. We will show how to perform this using StandardScalar.

In [45]:
```
from sklearn.preprocessing import StandardScaler

#Extract two real valued columns
housing_sub = housing[["housing_median_age","total_rooms"]]

scaler = StandardScaler() #initiate class
#Calling .fit lets scaler calculate the mean and standard deviation, i.e. trains the standardizer
scaler.fit(housing_sub)
print("Mean: ",scaler.mean_)
print("Std: ",scaler.scale_)

#To perform the standardization, use the .transform function
housing_std= scaler.transform(housing_sub)
print("Transfrom output")
print(housing_std)

#As a shorthand, the function .fit_transform performs both operations
housing_std_2= scaler.fit_transform(housing_sub)
print("Fit Transfrom output")
print(housing_std_2)
```

```
Mean:  [  28.63948643 2635.7630814 ]
Std:  [  12.58525273 2181.56240174]
Transfrom output
[[ 0.98214266 -0.8048191 ]
 [-0.60701891  2.0458901 ]
 [ 1.85618152 -0.53574589]
 ...
 [-0.92485123 -0.17499526]
 [-0.84539315 -0.35559977]
 [-1.00430931  0.06840827]]
Fit Transfrom output
[[ 0.98214266 -0.8048191 ]
 [-0.60701891  2.0458901 ]
 [ 1.85618152 -0.53574589]
 ...
 [-0.92485123 -0.17499526]
 [-0.84539315 -0.35559977]
 [-1.00430931  0.06840827]]
```

Prepare Data using a pipeline

Now, we will show how we can use scikit learn to create a pipeline that performs all the data preparation in one clean function call. For simplicity, we will not perform the closest city feature extraction in this pipeline.

It is very useful to combine several steps into one to make the process much simpler to understand and easy to alter.

In [46]:
```
housing = load_housing_data(DATASET_PATH) # Load the dataset

housing_features = housing.drop("median_house_value", axis=1) # drop labels for training set features
                                                               # the input to the model should not contain the true label
housing_target = housing["median_house_value"].copy()
```

In [47]:
```
housing_features.head()
```

Out[47]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	NEAR BAY

In [48]:
```
# This cell implements the complete pipeline for preparing the data
# using sklearns TransformerMixins
# Earlier we mentioned different types of features: categorical, and floats.
# In the case of floats we might want to convert them to categories.
# On the other hand categories in which are not already represented as integers must be mapped to integers before
# feeding to the model.

# Additionally, categorical values could either be represented as one-hot vectors or simple as normalized/unnormali
# Here we encode them using one hot vectors.

# DO NOT WORRY IF YOU DO NOT UNDERSTAND ALL THE STEPS OF THIS PIPELINE. CONCEPTS LIKE NORMALIZATION,
# ONE-HOT ENCODING ETC. WILL ALL BE COVERED IN DISCUSSION

from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OneHotEncoder

from sklearn.base import BaseEstimator, TransformerMixin
, →

######Processing Real Valued Features
# column indices
rooms_ix, bedrooms_ix, population_ix, households_ix = 3, 4, 5, 6

class AugmentFeatures(BaseEstimator, TransformerMixin):
    '''
    implements the previous features we had defined
    housing["rooms_per_household"] = housing["total_rooms"]/housing["households"]
    housing["bedrooms_per_room"] = housing["total_bedrooms"]/housing["total_rooms"]
    housing["population_per_household"]=housing["population"]/housing["households"]
    '''
    def __init__(self, add_bedrooms_per_room = True):
        self.add_bedrooms_per_room = add_bedrooms_per_room
    def fit(self, X, y=None):
        return self  # nothing else to do
    def transform(self, X):
        #Note that we do not use the pandas indexing anymore
        #This is due to sklearn transforming the dataframe into a numpy array during the processing
        #Thus, depending on where AugmentFeatures is in the pipeline, a different input type can be expected
        rooms_per_household = X[:, rooms_ix] / X[:, households_ix]
        population_per_household = X[:, population_ix] / X[:, households_ix]
        if self.add_bedrooms_per_room:
            bedrooms_per_room = X[:, bedrooms_ix] / X[:, rooms_ix]
            return np.c_[X, rooms_per_household, population_per_household,
                         bedrooms_per_room]
        else:
            return np.c_[X, rooms_per_household, population_per_household]

#Example of using AugmentFeatures
housing_features_num = housing_features.drop("ocean_proximity", axis=1) # remove the categorical features
attr_adder = AugmentFeatures(add_bedrooms_per_room=False) #Create transformer object
housing_extra_attribs = attr_adder.transform(housing_features_num.values) #housing_num.values extracts the numpy ar

print("Example of Augment Features Transformer")
print(housing_extra_attribs[0])


#Pipiline for real valued features
num_pipeline = Pipeline([
        ('imputer', SimpleImputer(strategy="median")), #Imputes using median
        ('attribs_adder', AugmentFeatures(add_bedrooms_per_room=True)), #
        ('std_scaler', StandardScaler()),
    ])



#Example
#Output is a numpy array
housing_features_num_tr = num_pipeline.fit_transform(housing_features_num)
print("Example Output of Pipeline for numerical output")
print(housing_features_num_tr[0])
```

```
Example of Augment Features Transformer
[-122.23         37.88         41.          880.          129.
  322.          126.            8.3252        6.98412698    2.55555556]
Example Output of Pipeline for numerical output
[-1.32783522  1.05254828  0.98214266 -0.8048191  -0.97247648 -0.9744286
 -0.97703285  2.34476576  0.62855945 -0.04959654 -1.02998783]
```

In [49]:
```
#Full Pipeline

#Splits names into numerical and categorical features
numerical_features = list(housing_features_num)
categorical_features = ["ocean_proximity"]

#Applies different transformations on numerical columns vs categorial columns
full_pipeline = ColumnTransformer([
        ("num", num_pipeline, numerical_features),
        ("cat", OneHotEncoder(), categorical_features),
    ])


#Example of full pipeline
#Output is a numpy array
housing_prepared = full_pipeline.fit_transform(housing_features)
print("Example Output of full Pipeline")
print(housing_prepared[0])
```

```
Example Output of full Pipeline
[-1.32783522  1.05254828  0.98214266 -0.8048191  -0.97247648 -0.9744286
 -0.97703285  2.34476576  0.62855945 -0.04959654 -1.02998783  0.
  0.          0.          1.          0.        ]
```

Now, we have a pipeline that easily processes the input data into our desired form.

Splitting our dataset

First we need to carve out our dataset into a training and testing cohort. To do this we'll use train_test_split, a very elementary tool that arbitrarily splits the data into training and testing cohorts.

Note that we first perform the train test split on the data before it was processed in the pipeline and then separatelyprocess the train and test data. This is done to avoid injecting information into the test data from the train data such filling in missing values in the test data with knowledge of the train data.

In [50]:
```
from sklearn.model_selection import train_test_split
data_target = housing['median_house_value']
train, test, target, target_test = train_test_split(housing_features, data_target, test_size=0.3, random_state=0)

train = full_pipeline.fit_transform(train)
test = full_pipeline.fit_transform(test)
```

Select a model and train

Once we have prepared the dataset it's time to choose a model.

As our task is to predict the median_house_value (a floating value), regression is well suited for this.

In [51]:
```
from sklearn.linear_model import LinearRegression

#Instantiate a linear regresion class
lin_reg = LinearRegression()
#Train the class using the .fit function
lin_reg.fit(train, target)

# let's try the full preprocessing pipeline on a few training instances
data = test
labels = target_test

#Uses predict to get the predicted target values
print("Predictions:", lin_reg.predict(data)[:5])
print("Actual labels:", list(labels)[:5])
```

```
Predictions: [210975.9892164  283834.89185828 179131.95542365  92162.26714094
 295068.95402291]
Actual labels: [136900.0, 241300.0, 200700.0, 72500.0, 460000.0]
```

In [52]:
```
from sklearn.metrics import mean_squared_error

preds = lin_reg.predict(test)
mse = mean_squared_error(target_test, preds)
rmse = np.sqrt(mse)
rmse
```

Out[52]:
```
69145.58671722481
```

TODO: Applying the end-end ML steps to a different dataset.

We will apply what we've learnt to another dataset (NYC airbnb dataset from 2019 (https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data)). We will predict airbnb price based on other features.

Note: You do not have to use only one cell when programming your code and can do it over multiple cells.

[50 pts] Visualizing Data

[10 pts] Load the data + statistics

- Load the dataset: airbnb/AB_NYC_2019.csv and display the first 5 few rows of the data

In [53]:
```
#Your code
DATASET_PATH = os.path.join("datasets", "airbnb")
csv_path = os.path.join(DATASET_PATH, "AB_NYC_2019.csv")
airbnb = pd.read_csv(csv_path)
airbnb.head()
```

Out[53]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_re
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

- Pull up info on the data type for each of the data fields. Will any of these be problematic feeding into your model (you may need to do a little research on this)? Discuss:

In [54]:
```
#Your code
airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

The name object is a string that cannot be one hot encoded. The id is a unique identification value with no quantitative or qualitative value. The host name and id are dropped for the same reason. The last review is simply a date with no quant/qual value. Im not sure why we drop reviews per month as if a place is reviewed many times a month it could indicate that the place is cheap.

- Drop the following columns: name, id, host_id, host_name, last_review, and reviews_per_month and display first 5 rows

In [55]:
```
#Your code
airbnb = airbnb.drop(['name', 'id', 'host_id', 'host_name', 'last_review', 'reviews_per_month', 'neighbourhood'], a
airbnb.head()
```

Out[55]:

	neighbourhood_group	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
0	Brooklyn	40.64749	-73.97237	Private room	149	1	9	6	365
1	Manhattan	40.75362	-73.98377	Entire home/apt	225	1	45	2	355
2	Manhattan	40.80902	-73.94190	Private room	150	3	0	1	365
3	Brooklyn	40.68514	-73.95976	Entire home/apt	89	1	270	1	194
4	Manhattan	40.79851	-73.94399	Entire home/apt	80	10	9	1	0

- Display a summary of the statistics of the loaded data using .describe

In [56]:
```
#Your code
airbnb.describe()
```

Out[56]:

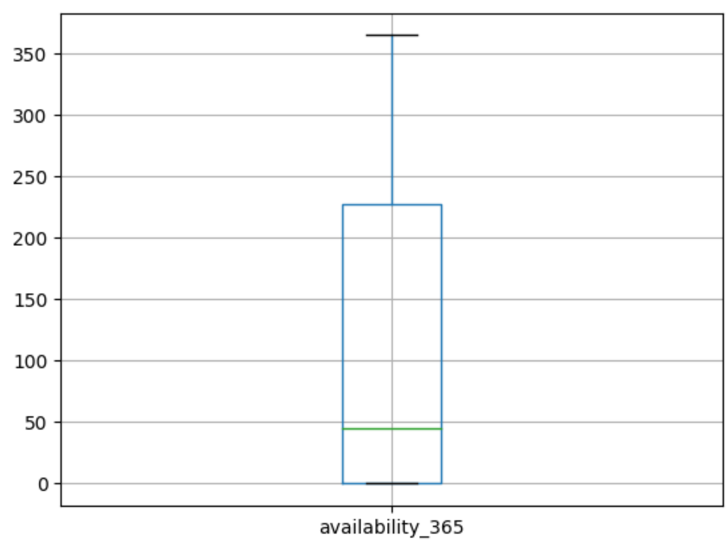
	latitude	longitude	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
count	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000
mean	40.728949	-73.952170	152.720687	7.029962	23.274466	7.143982	112.781327
std	0.054530	0.046157	240.154170	20.510550	44.550582	32.952519	131.622289
min	40.499790	-74.244420	0.000000	1.000000	0.000000	1.000000	0.000000
25%	40.690100	-73.983070	69.000000	1.000000	1.000000	1.000000	0.000000
50%	40.723070	-73.955680	106.000000	3.000000	5.000000	1.000000	45.000000
75%	40.763115	-73.936275	175.000000	5.000000	24.000000	2.000000	227.000000
max	40.913060	-73.712990	10000.000000	1250.000000	629.000000	327.000000	365.000000

[10 pts] Plot [boxplots (https://en.wikipedia.org/wiki/Box_plot)](https://en.wikipedia.org/wiki/Box_plot) for the following 3 features: availability_365, number_of_reviews, price

You may use either pandas or matplotlib to plot the boxplot

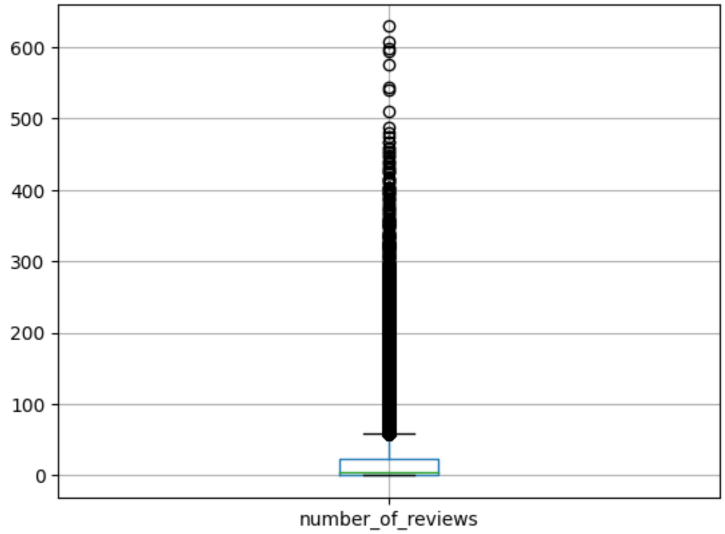
In [57]:
```
#Your code
airbnb.boxplot(column=['availability_365'])
```

Out[57]:
```
<AxesSubplot: >
```



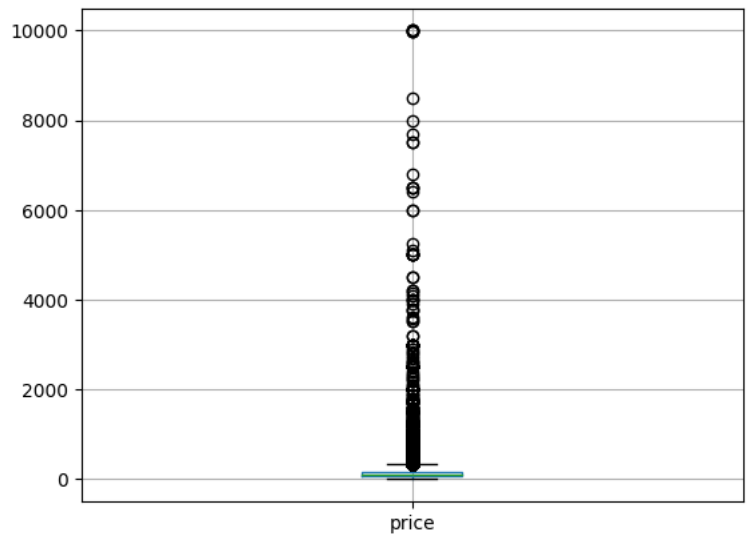
In [58]:
```
airbnb.boxplot(column=['number_of_reviews'])
```

Out[58]: <AxesSubplot: >



In [59]:
```
airbnb.boxplot(column=['price'])
```

Out[59]: <AxesSubplot: >



- What do you observe from the boxplot about the features? Anything suprising?

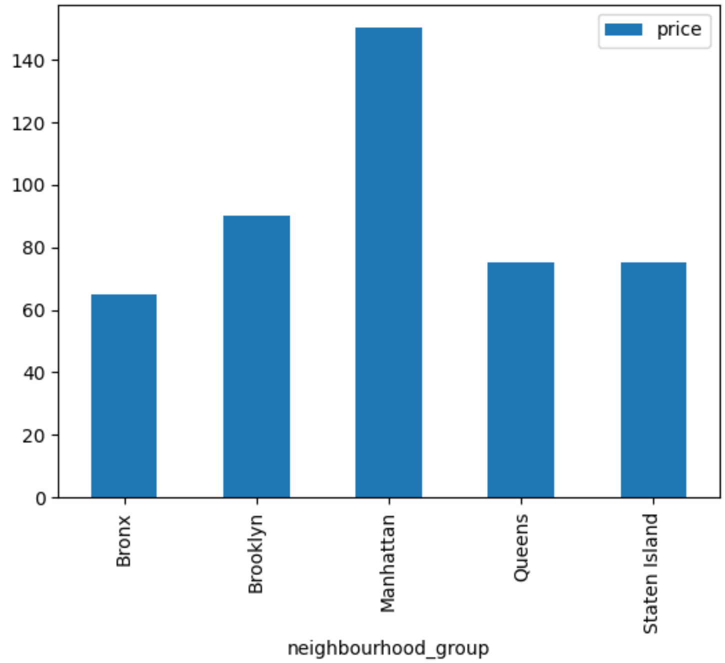
Most rooms are available just under 50 days out of the year. The average number of reviews on properties are quite low. The average price of a listing is also very low, but most outliers get to about 3.5k

[10 pts] Plot median price of a listing per neighbourhood_group using a bar plot

In [60]:
```
#Your code
airbnb_neighborhood_group_median = airbnb.groupby("neighbourhood_group").median()

airbnb_neighborhood_group_median.plot.bar(y ="price")
```

```
/tmp/ipykernel_10764/1367792229.py:2: FutureWarning: The default value of numeric_only in DataFrameGroupBy.median
is deprecated. In a future version, numeric_only will default to False. Either specify numeric_only or select only
columns which should be valid for the function.
  airbnb_neighborhood_group_median = airbnb.groupby("neighbourhood_group").median()
```

Out[60]: <AxesSubplot: xlabel='neighbourhood_group'>



- Describe what you expected to see with these features and what you actually observed

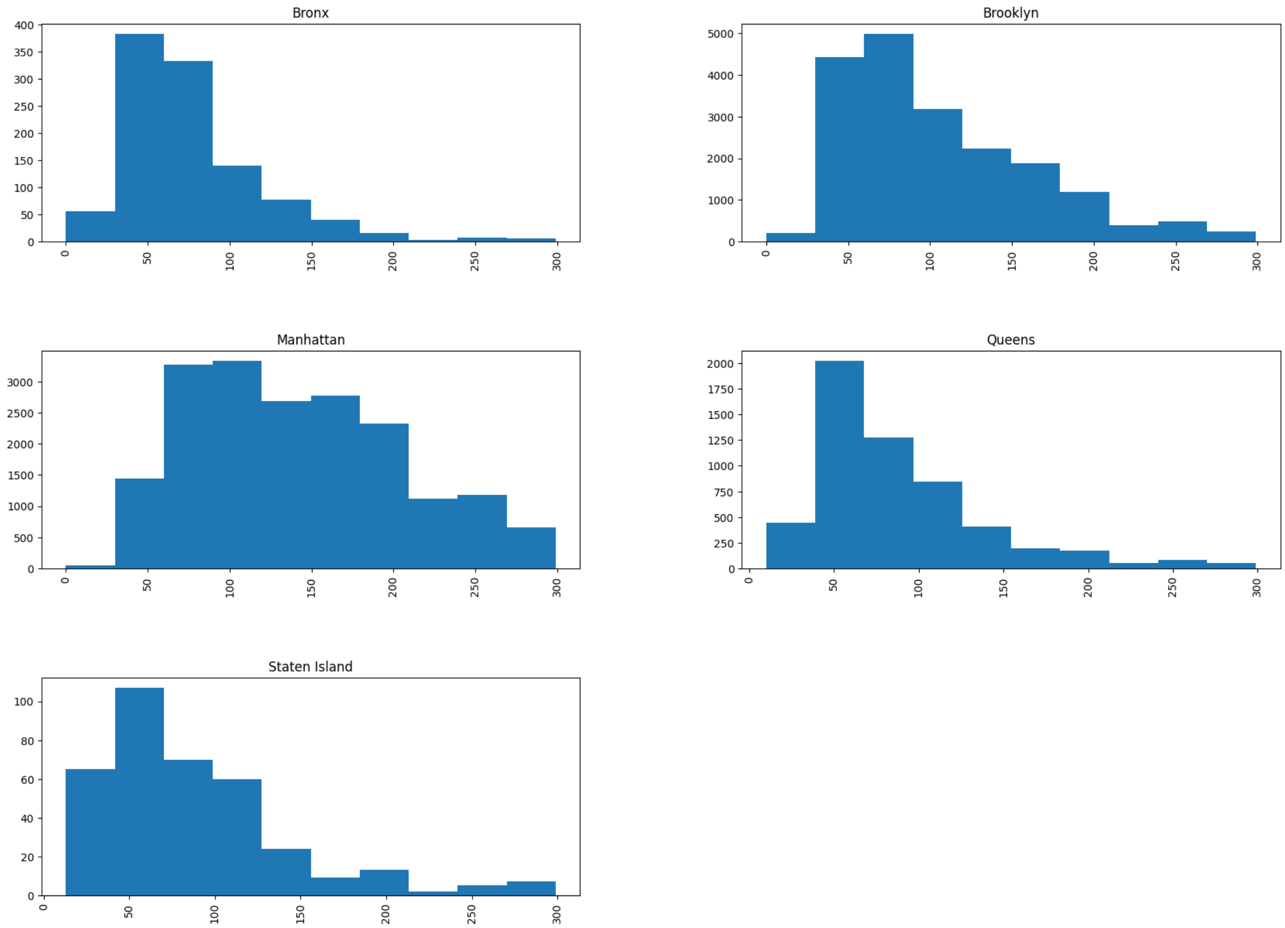
Do not know much about the specific neighborhood of new york but I did expect some neighborhoods to be more expensive than others.

- So we can see different neighborhoods have dramatically different pricepoints, but how does the price breakdown by range. To see let's do a histogram of price by neighborhood to get a better sense of the distribution.

To prevent outliers from affecting the histogram, use the input *range = [0,300]* in the histogram function which will upperbound the max price to 300 and ignore the outliers.

In [61]:
```
#Your code
airbnb_cheap= airbnb[(airbnb['price'] < 300)]
#airbnb_expensive.head()
airbnb_cheap["price"].hist(by= airbnb_cheap["neighbourhood_group"],figsize=(20,15))
```

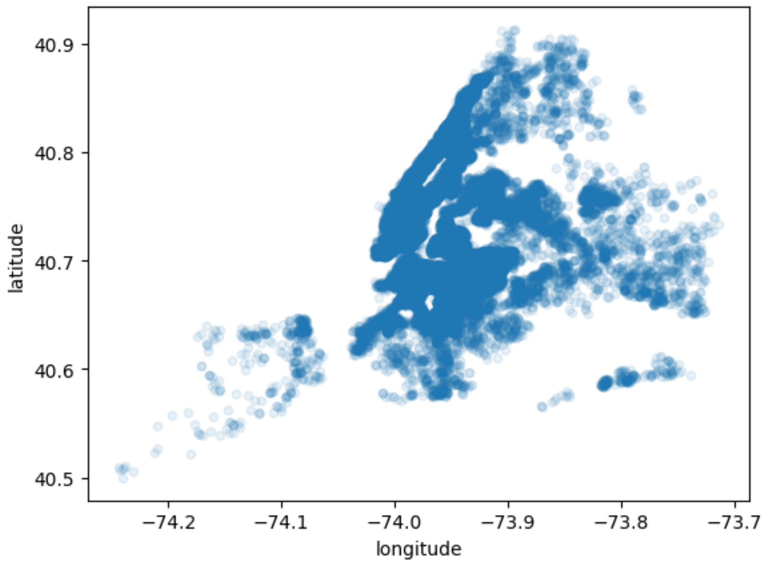
Out[61]:
```
array([[<AxesSubplot: title={'center': 'Bronx'}>,
        <AxesSubplot: title={'center': 'Brooklyn'}>],
       [<AxesSubplot: title={'center': 'Manhattan'}>,
        <AxesSubplot: title={'center': 'Queens'}>],
       [<AxesSubplot: title={'center': 'Staten Island'}>,
        <AxesSubplot: >]], dtype=object)
```



[5 pts] Plot a map of airbnbs throughout New York. You do not need to overlay a map.

In [62]:
```
#Your code
airbnb_cheap.plot(kind="scatter", x="longitude", y="latitude", alpha=.1)
```

Out[62]: <AxesSubplot: xlabel='longitude', ylabel='latitude'>

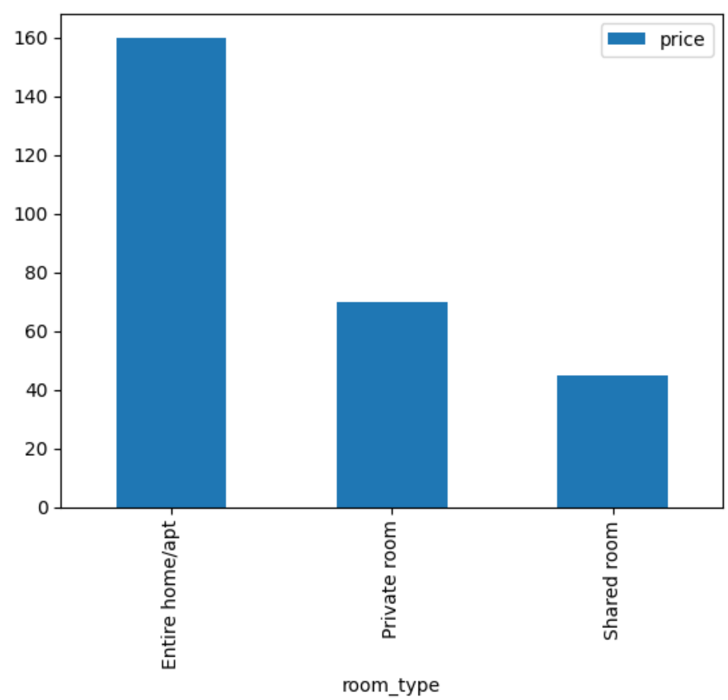


[10 pts] Plot median price of room types who have availability greater than 180 days and neighbourhood_group is Manhattan

In [63]:
```
#Your code
airbnb_stuff= airbnb[(airbnb['availability_365'] > 180) & (airbnb["neighbourhood_group"] == "Manhattan")]
airbnb_stuff= airbnb.groupby("room_type").median()
airbnb_stuff.plot.bar(y ="price")
```

```
/tmp/ipykernel_10764/3518313448.py:3: FutureWarning: The default value of numeric_only in DataFrameGroupBy.median
is deprecated. In a future version, numeric_only will default to False. Either specify numeric_only or select only
columns which should be valid for the function.
  airbnb_stuff= airbnb.groupby("room_type").median()
```

Out[63]: <AxesSubplot: xlabel='room_type'>



[5 pts] Find features that correlate with price

Using the correlation matrix:

- which features have positive correlation with the price?
- which features have negative correlation with the price?

In [64]:
```
#Your code
corr_matrix = airbnb.corr()
corr_matrix["price"].sort_values(ascending=False)
```

```
/tmp/ipykernel_10764/3060378573.py:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprec
ated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_onl
y to silence this warning.
  corr_matrix = airbnb.corr()
```

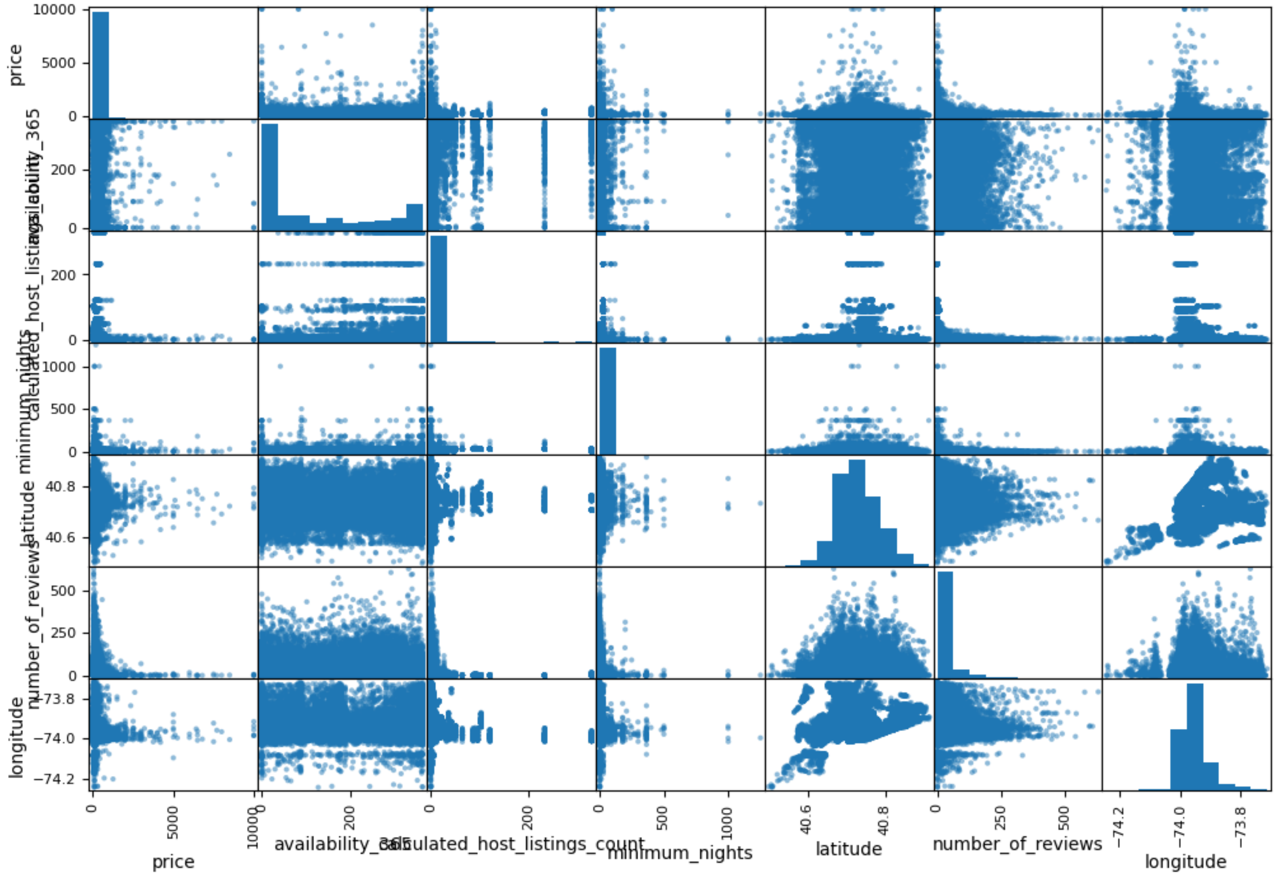
Out[64]:
```
price                             1.000000
availability_365                  0.081829
calculated_host_listings_count    0.057472
minimum_nights                    0.042799
latitude                          0.033939
number_of_reviews                -0.047954
longitude                        -0.150019
Name: price, dtype: float64
```

Availability, host listing counts, minimum nights, and latitude have a positive correlation with price. Number of reviews and longitude have a negative correlation with price

- Plot the full Scatter Matrix to see the correlation between prices and the other features

In [65]:
```
#Your code
attributes = ["price", "availability_365", "calculated_host_listings_count", "minimum_nights", "latitude",
              "number_of_reviews", "longitude"]
scatter_matrix(airbnb[attributes], figsize=(12, 8))
```

Out[65]:
```
array([[<AxesSubplot: xlabel='price', ylabel='price'>,
        <AxesSubplot: xlabel='availability_365', ylabel='price'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='price'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='price'>,
        <AxesSubplot: xlabel='latitude', ylabel='price'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='price'>,
        <AxesSubplot: xlabel='longitude', ylabel='price'>],
       [<AxesSubplot: xlabel='price', ylabel='availability_365'>,
        <AxesSubplot: xlabel='availability_365', ylabel='availability_365'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='availability_365'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='availability_365'>,
        <AxesSubplot: xlabel='latitude', ylabel='availability_365'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='availability_365'>,
        <AxesSubplot: xlabel='longitude', ylabel='availability_365'>],
       [<AxesSubplot: xlabel='price', ylabel='calculated_host_listings_count'>,
        <AxesSubplot: xlabel='availability_365', ylabel='calculated_host_listings_count'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='calculated_host_listings_count'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='calculated_host_listings_count'>,
        <AxesSubplot: xlabel='latitude', ylabel='calculated_host_listings_count'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='calculated_host_listings_count'>,
        <AxesSubplot: xlabel='longitude', ylabel='calculated_host_listings_count'>],
       [<AxesSubplot: xlabel='price', ylabel='minimum_nights'>,
        <AxesSubplot: xlabel='availability_365', ylabel='minimum_nights'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='minimum_nights'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='minimum_nights'>,
        <AxesSubplot: xlabel='latitude', ylabel='minimum_nights'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='minimum_nights'>,
        <AxesSubplot: xlabel='longitude', ylabel='minimum_nights'>],
       [<AxesSubplot: xlabel='price', ylabel='latitude'>,
        <AxesSubplot: xlabel='availability_365', ylabel='latitude'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='latitude'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='latitude'>,
        <AxesSubplot: xlabel='latitude', ylabel='latitude'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='latitude'>,
        <AxesSubplot: xlabel='longitude', ylabel='latitude'>],
       [<AxesSubplot: xlabel='price', ylabel='number_of_reviews'>,
        <AxesSubplot: xlabel='availability_365', ylabel='number_of_reviews'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='number_of_reviews'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='number_of_reviews'>,
        <AxesSubplot: xlabel='latitude', ylabel='number_of_reviews'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='number_of_reviews'>,
        <AxesSubplot: xlabel='longitude', ylabel='number_of_reviews'>],
       [<AxesSubplot: xlabel='price', ylabel='longitude'>,
        <AxesSubplot: xlabel='availability_365', ylabel='longitude'>,
        <AxesSubplot: xlabel='calculated_host_listings_count', ylabel='longitude'>,
        <AxesSubplot: xlabel='minimum_nights', ylabel='longitude'>,
        <AxesSubplot: xlabel='latitude', ylabel='longitude'>,
        <AxesSubplot: xlabel='number_of_reviews', ylabel='longitude'>,
        <AxesSubplot: xlabel='longitude', ylabel='longitude'>]],
      dtype=object)
```



[30 pts] Prepare the Data

[5 pts] Partition the data into the features and the target data. The target data is price. Then partition the feature data into categorical and numerical features.

In [66]:
```
#Your code
airbnb_features = airbnb.drop(["price"], axis=1)
data_target = airbnb['price']
train, test, target, target_test = train_test_split(airbnb_features, data_target, test_size=0.3, random_state=0)
train_num = train.drop(['neighbourhood_group', 'room_type'], axis=1)
test_num = test.drop(['neighbourhood_group', 'room_type'], axis=1)
train_cat = train.drop(["availability_365", "calculated_host_listings_count", "minimum_nights", "latitude",
             "number_of_reviews", "longitude"], axis=1)
test_cat = test.drop(["availability_365", "calculated_host_listings_count", "minimum_nights", "latitude",
             "number_of_reviews", "longitude"], axis=1)
train_num.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34226 entries, 13115 to 2732
Data columns (total 6 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   latitude                        34226 non-null  float64
 1   longitude                       34226 non-null  float64
 2   minimum_nights                  34226 non-null  int64
 3   number_of_reviews               34226 non-null  int64
 4   calculated_host_listings_count  34226 non-null  int64
 5   availability_365                34226 non-null  int64
dtypes: float64(2), int64(4)
memory usage: 1.8 MB
```

[10 pts] Create a scikit learn Transformer that augments the numerical data with the following two features

- Max_yearly_bookings = availability_365 / minimum_nights
- Distance from airbnb to the NYC JFK Airport
  - Latitude: 40.641766 , Longitude: -73.780968

Make sure to append these new features in this order.

You may use the previously defined distance_func for the distance calculation.

Note that this Transformer will be applied after imputation so we do not have to worry about Nulls in the data.

In [67]:
```
######Processing Real Valued Features
# column indices
lat_ix, long_ix, availability_ix, min_nights_ix = 0, 1, 5, 2

class AugmentFeatures(BaseEstimator, TransformerMixin):
    '''
    implements max yearly bookings and distance from airport
    airbnb["max_yearly_bookings"] = airbnb["availability_365"]/housing["minimum_nights"]
    airbnb["distance_to_airport"] = distance_func((airbnb["latitude"], airbnb["longitude"],(40.641766, -73.780968))
    '''
    def __init__(self, augment = True):
        self.augment = augment
    def fit(self, X, y=None):
        return self  # nothing else to do
    def transform(self, X):
        #Note that we do not use the pandas indexing anymore
        #This is due to sklearn transforming the dataframe into a numpy array during the processing
        #Thus, depending on where AugmentFeatures is in the pipeline, a different input type can be expected
        max_yearly_bookings = X[:, availability_ix] / X[:, min_nights_ix]
        dist_to_airport = distance_func((X[:, lat_ix], X[:, long_ix]), (40.641766, -73.780968))
        if self.augment:
            return np.c_[X, max_yearly_bookings, dist_to_airport]
        else:
            return np.c_[X]
```

-Test your new agumentation class by applying it to the numerical data you created. Print out the first 3 rows of the resultant data.

Do not worry about missing data since none of the features we used involved nulls.

In [68]:
```
#Your code
#Example of using AugmentFeatures
attr_adder = AugmentFeatures(augment=True) #Create transformer object
train_extra = attr_adder.transform(train_num.values) #housing_num.values extracts the numpy array of the dataframe
attr_adder = AugmentFeatures(augment=True) #Create transformer object
test_extra = attr_adder.transform(test_num.values) #housing_num.values extracts the numpy array of the dataframe

print(train_extra[:3])
print(test_extra[:3])
```

```
[[ 40.71569 -73.93735   7.         0.         1.         0.         0.
   15.54    ]
 [ 40.76222 -73.99088   3.         4.         1.         0.         0.
   22.19    ]
 [ 40.61922 -73.99399   1.        24.         1.       183.       183.
   18.15    ]]
[[ 40.7243  -74.0111    3.         0.         1.        42.        14.
   21.47    ]
 [ 40.72555 -73.99283   1.         5.         1.        75.        75.
   20.15    ]
 [ 40.71687 -73.95012   5.         5.         3.        31.         6.2
   16.53    ]]
```

[10 pts] Create a sklearn pipeline that performs the following operations of the feature data

Now, we will create a full pipeline that processes the data before creating the model.

For the numerical data, perfrom the following operations in order:

- Use a SimpleImputer that imputes using the median value
- Use the custom feature augmentation made in the previous part
- Use StandardScaler to standardize the mean and standard deviation

For categorical features, perform the following:

- Perform one hot encoding on all the remaining categorical features: {neighbourhood_group, room_type}

After making the pipeline, perform the transform operation on the feature data and print out the first 3 rows.

In [69]:
```
#Your code
#Pipiline for real valued features
num_pipeline = Pipeline([
        ('imputer', SimpleImputer(strategy="median")), #Imputes using median
        ('attribs_adder', AugmentFeatures(augment=True)), #
        ('std_scaler', StandardScaler()),
    ])

#Splits names into numerical and categorical features
airbnb_features_num = airbnb_features.drop(['neighbourhood_group', 'room_type'], axis=1)
numerical_features = list(airbnb_features_num)
categorical_features = ['neighbourhood_group', 'room_type']

#Applies different transformations on numerical columns vs categorial columns
full_pipeline = ColumnTransformer([
        ("num", num_pipeline, numerical_features),
        ("cat", OneHotEncoder(), categorical_features),
    ])


#Example of full pipeline
#Output is a numpy array
airbnb_prepared = full_pipeline.fit_transform(airbnb_features)
print("Example Output of full Pipeline")
print(housing_prepared[:3])
```

```
Example Output of full Pipeline
[[-1.32783522  1.05254828  0.98214266 -0.8048191  -0.97247648 -0.9744286
  -0.97703285  2.34476576  0.62855945 -0.04959654 -1.02998783  0.
   0.          0.          1.          0.        ]
 [-1.32284391  1.04318455 -0.60701891  2.0458901   1.35714343  0.86143887
   1.66996103  2.33223796  0.32704136 -0.09251223 -0.8888972   0.
   0.          0.          1.          0.        ]
 [-1.33282653  1.03850269  1.85618152 -0.53574589 -0.82702426 -0.82077735
  -0.84363692  1.7826994   1.15562047 -0.02584253 -1.29168566  0.
   0.          0.          1.          0.        ]]
```

[5 pts] Set aside 20% of the data as test test (80% train, 20% test). Apply previously created pipeline to the train and test data separately as shown in the introduction example.

In [70]:
```
#Your code
data_target = airbnb['price']
train, test, target, target_test = train_test_split(airbnb_features, data_target, test_size=0.2, random_state=0)

train = full_pipeline.fit_transform(train)
test = full_pipeline.fit_transform(test)
```

[20 pts] Fit a Linear Regression Model

The task is to predict the price, you could refer to the housing example on how to train and evaluate your model using the mean squared error (MSE). Provide both test and train set MSE values.

In [71]:
```
#Your codes
#Instantiate a linear regresion class
lin_reg = LinearRegression()
#Train the class using the .fit function
lin_reg.fit(train, target)

# let's try the full preprocessing pipeline on a few training instances
data = test
labels = target_test

#Uses predict to get the predicted target values
#print("Predictions:", lin_reg.predict(data)[:5])
#print("Actual labels:", list(labels)[:5])

train_preds = lin_reg.predict(train)
mse = mean_squared_error(target, train_preds)
print("training MSE: " + str(mse))
test_preds = lin_reg.predict(test)
mse = mean_squared_error(target_test, test_preds)
print("test MSE: " + str(mse))
```

```
training MSE: 52635.8112639203
test MSE: 48606.52419328741
```