In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df = pd.read_csv("housing2.csv")
```

In [13]:
```
df.head()
```

Out[13]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households
0	-122.23	37.88	41.0	880	129.0	322.0	126
1	-122.22	37.86	21.0	7099	1106.0	2401.0	1138
2	-122.24	37.85	52.0	1467	190.0	496.0	177
3	-122.25	37.85	52.0	1274	235.0	558.0	219
4	-122.25	37.85	NaN	1627	280.0	NaN	259

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20382 non-null  float64
 3   total_rooms         20640 non-null  int64  
 4   total_bedrooms      15758 non-null  float64
 5   population          20596 non-null  float64
 6   households          19335 non-null  object 
 7   median_income       17873 non-null  float64
 8   median_house_value  20640 non-null  int64  
 9   ocean_proximity     20640 non-null  object 
 10  gender              16620 non-null  object 
dtypes: float64(6), int64(2), object(3)
memory usage: 1.7+ MB
```

In [10]:
```
df.ocean_proximity.value_counts()
```

Out[10]:
```
<1H OCEAN     9136
INLAND        6551
NEAR OCEAN    2658
NEAR BAY      2290
ISLAND           5
Name: ocean_proximity, dtype: int64
```

In [12]:
```
df.households.head()
```

Out[12]:
```
0     126
1    1138
2     177
3     219
4     259
Name: households, dtype: object
```

In [20]:
```
df.households.value_counts()
```

Out[20]:
```
no      3080
282       47
375       46
380       45
306       45
         ...
2905       1
3832       1
1503       1
1410       1
1026       1
Name: households, Length: 1703, dtype: int64
```

In [26]:
```
df.households.replace('no',0,inplace=True)
```

In [27]:
```
df.households.value_counts()
```

Out[27]:
```
0       3080
282       47
375       46
380       45
306       45
         ...
3832       1
1125       1
1503       1
1462       1
1584       1
Name: households, Length: 1703, dtype: int64
```

In [32]:
```
df.households.astype(float)
```

Out[32]:
```
0         126.0
1        1138.0
2         177.0
3         219.0
4         259.0
          ...
20635     330.0
20636     114.0
20637     433.0
20638     349.0
20639     530.0
Name: households, Length: 20640, dtype: float64
```

In [34]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20382 non-null  float64
 3   total_rooms         20640 non-null  int64  
 4   total_bedrooms      15758 non-null  float64
 5   population          20596 non-null  float64
 6   households          19335 non-null  object 
 7   median_income       17873 non-null  float64
 8   median_house_value  20640 non-null  int64  
 9   ocean_proximity     20640 non-null  object 
 10  gender              16620 non-null  object 
dtypes: float64(6), int64(2), object(3)
memory usage: 1.7+ MB
```

In [36]:
```
df.households.isnull().sum()
```

Out[36]:
```
1305
```

In [39]:
```
df.households.fillna(0,inplace=True)
```

In [40]:
```
df.households.isnull().sum()
```

Out[40]:
```
0
```

In [54]:
```
df.households = df.households.astype(int)
```

In [55]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           20640 non-null  float64
 1   latitude            20640 non-null  float64
 2   housing_median_age  20382 non-null  float64
 3   total_rooms         20640 non-null  int64  
 4   total_bedrooms      15758 non-null  float64
 5   population          20596 non-null  float64
 6   households          20640 non-null  int32  
 7   median_income       17873 non-null  float64
 8   median_house_value  20640 non-null  int64  
 9   ocean_proximity     20640 non-null  object 
 10  gender              16620 non-null  object 
dtypes: float64(6), int32(1), int64(2), object(2)
memory usage: 1.7+ MB
```

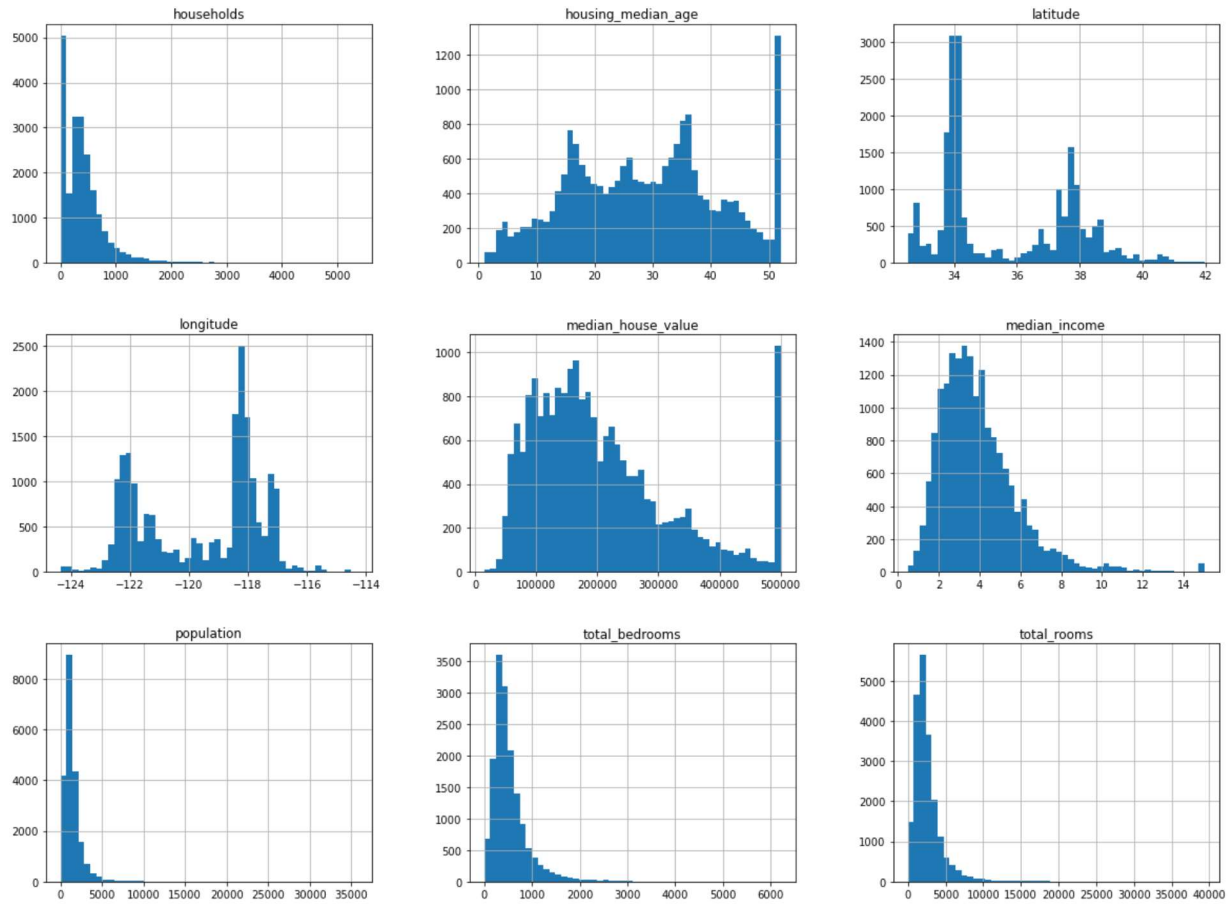
In [57]:
```
df.describe()
```

Out[57]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populati
count	20640.000000	20640.000000	20382.000000	20640.000000	15758.000000	20596.0000
mean	-119.569704	35.631861	28.676283	2635.763081	539.920104	1424.9287
std	2.003532	2.135952	12.589284	2181.615252	419.834171	1132.2377
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.0000
25%	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.0000
50%	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.0000
75%	-118.010000	37.710000	37.000000	3148.000000	652.000000	1725.0000
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.0000

In [58]:
```
%matplotlib inline
import matplotlib.pyplot as plt
```

In [59]:
```
df.hist(bins=50,figsize=(20,15))
plt.show()
```



In [60]:
```
df.isnull().sum()
```

Out[60]:
```
longitude              0
latitude               0
housing_median_age     258
total_rooms            0
total_bedrooms         4882
population             44
households             0
median_income          2767
median_house_value     0
ocean_proximity        0
gender                 4020
dtype: int64
```

In [67]:
```
df.housing_median_age.fillna(df.housing_median_age.mean(),inplace=True)
df.total_bedrooms.fillna(df.total_bedrooms.mean(),inplace=True)
df.population.fillna(df.population.mean(),inplace=True)
df.median_income.fillna(df.median_income.mean(),inplace=True)
```

In [68]:
```
df.isnull().sum()
```

Out[68]:
```
longitude               0
latitude                0
housing_median_age      0
total_rooms             0
total_bedrooms          0
population              0
households              0
median_income           0
median_house_value      0
ocean_proximity         0
gender               4020
dtype: int64
```

In [72]:
```
df.drop(['gender'],axis=1,inplace=True)
```

In [74]:
```
df.duplicated().sum()
```

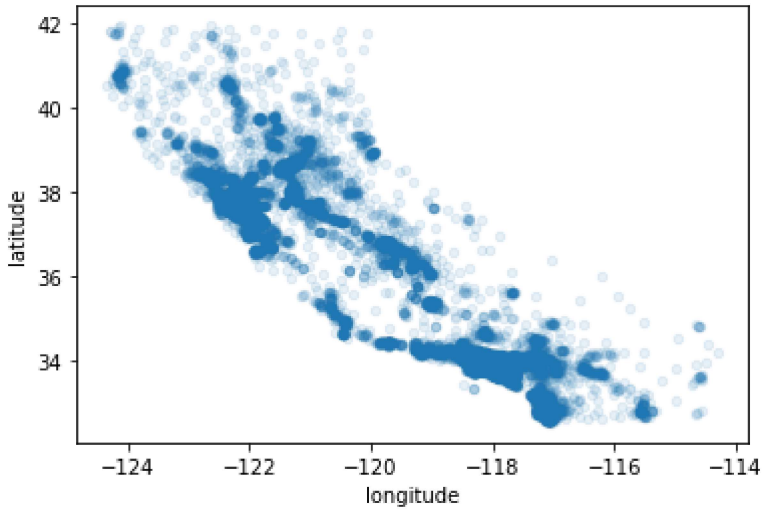
Out[74]:
```
0
```

In [75]:
```
df.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  int64  
 4   total_bedrooms      20640 non-null  float64
 5   population          20640 non-null  float64
 6   households          20640 non-null  int32  
 7   median_income       20640 non-null  float64
 8   median_house_value  20640 non-null  int64  
 9   ocean_proximity     20640 non-null  object 
dtypes: float64(6), int32(1), int64(2), object(1)
memory usage: 1.5+ MB
```

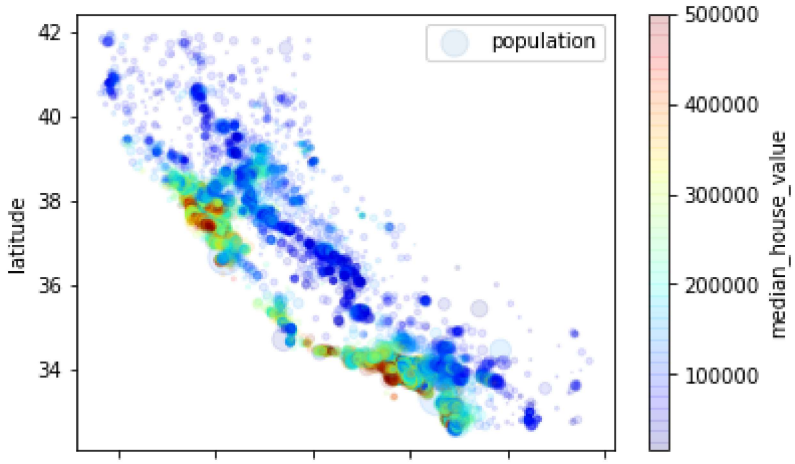
In [76]:
```
df.plot(kind="scatter",x='longitude',y='latitude',alpha=0.1)
```

Out[76]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f9bda51ca0>
```



In [77]:
```
ude',y='latitude',alpha=0.1,s=df.population/100,label="population",c="median_hous
```

Out[77]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1f9bdad1190>
```



In [78]:
```
corr_matrix = df.corr()
corr_matrix
```

Out[78]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	popu
longitude	1.000000	-0.924664	-0.106884	0.044568	0.063468	0.1
latitude	-0.924664	1.000000	0.009689	-0.036100	-0.054250	-0.1
housing_median_age	-0.106884	0.009689	1.000000	-0.356480	-0.296786	-0.2
total_rooms	0.044568	-0.036100	-0.356480	1.000000	0.793059	0.8
total_bedrooms	0.063468	-0.054250	-0.296786	0.793059	1.000000	0.7
population	0.100253	-0.109120	-0.291137	0.856124	0.743033	1.0
households	-0.010329	-0.005938	-0.160518	0.662329	0.809705	0.6
median_income	0.010336	-0.094187	-0.107553	0.189197	-0.008316	0.0
median_house_value	-0.045967	-0.144160	0.106648	0.134153	0.044949	-0.0

In [80]:
```
corr_matrix.median_house_value.sort_values()
```

Out[80]:
```
latitude            -0.144160
longitude           -0.045967
population          -0.024351
households           0.035346
total_bedrooms       0.044949
housing_median_age   0.106648
total_rooms          0.134153
median_income        0.650304
median_house_value   1.000000
Name: median_house_value, dtype: float64
```

In [85]:
```
df_info = df.describe()
```

In [92]:
```
df_info.loc['median'] = df.median()
df_info.loc['skew'] = df.skew()
df_info.loc['kurtosis']=df.kurt()
```

In [93]:
```
df_info
```

Out[93]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	popul
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.00
mean	-119.569704	35.631861	28.676283	2635.763081	539.920104	1424.92
std	2.003532	2.135952	12.510350	2181.615252	366.834544	1131.03
min	-124.350000	32.540000	1.000000	2.000000	1.000000	3.00
25%	-121.800000	33.930000	18.000000	1447.750000	338.000000	788.00
50%	-118.490000	34.260000	29.000000	2127.000000	539.920104	1167.00
75%	-118.010000	37.710000	37.000000	3148.000000	566.000000	1723.00
max	-114.310000	41.950000	52.000000	39320.000000	6210.000000	35682.00
median	-118.490000	34.260000	29.000000	2127.000000	539.920104	1167.00
skew	-0.297801	0.465953	0.057274	4.147343	3.753456	4.95
kurtosis	-1.330152	-1.117760	-0.775939	32.630927	26.430339	73.93

In [91]:
```

```

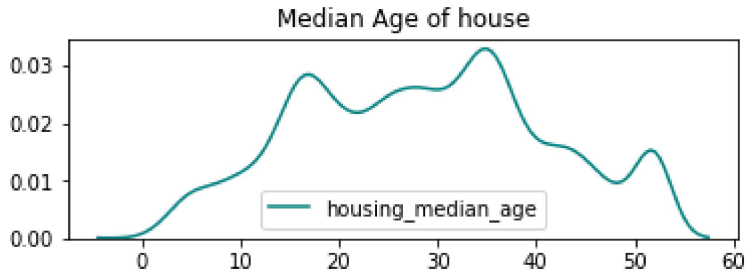
Out[91]:

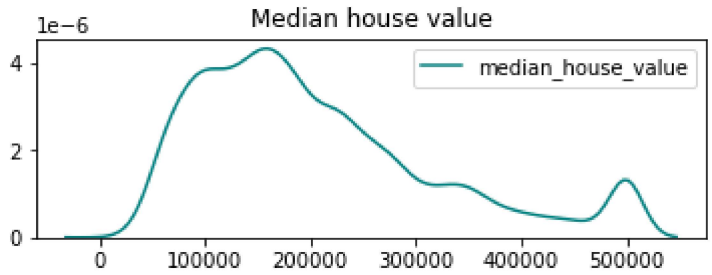
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	househ
0	-122.23	37.88	41.000000	880.0	129.0	322.000000	1
1	-122.22	37.86	21.000000	7099.0	1106.0	2401.000000	11
2	-122.24	37.85	52.000000	1467.0	190.0	496.000000	1
3	-122.25	37.85	52.000000	1274.0	235.0	558.000000	2
4	-122.25	37.85	28.676283	1627.0	280.0	1424.928724	2
...	...	...	...	...	...	...	
20635	-121.09	39.48	25.000000	1665.0	374.0	845.000000	3
20636	-121.21	39.49	18.000000	697.0	150.0	356.000000	1
20637	-121.22	39.43	17.000000	2254.0	485.0	1007.000000	4
20638	-121.32	39.43	18.000000	1860.0	409.0	741.000000	3
20639	-121.24	39.37	16.000000	2785.0	616.0	1387.000000	5

20640 rows × 10 columns

Multi Model Distribution

In [94]:
```
plt.subplot(2,1,1)
plt.title('Median Age of house')
sns.kdeplot(df.housing_median_age,color='teal')
plt.show()
plt.subplot(2,1,2)
plt.title('Median house value')
sns.kdeplot(df.median_house_value,color='teal')
plt.show()
```

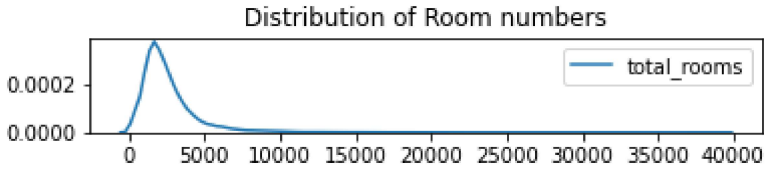


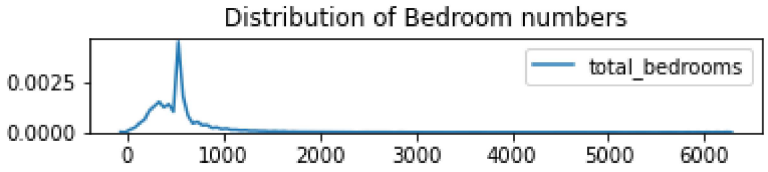


Almost Normal Distribution for total Rooms & population

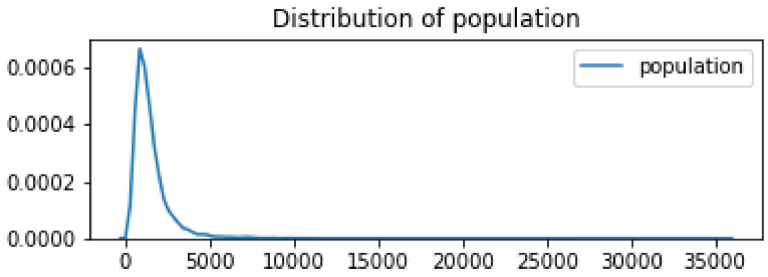
Total Bedroom & Householders due to cleaning data and fill NA with mean in Bedroom and Zero in householders (It isn't the perfect way for handling NA)

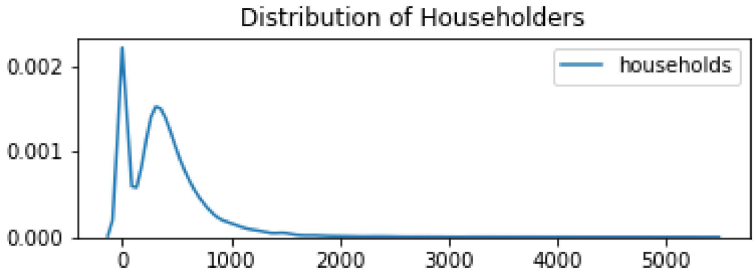
In [110]:
```
plt.subplot(4,1,1)
plt.title("Distribution of Room numbers")
sns.kdeplot(df.total_rooms)
plt.show()
plt.subplot(4,1,2)
plt.title("Distribution of Bedroom numbers")
sns.kdeplot(df.total_bedrooms)
plt.show()
```



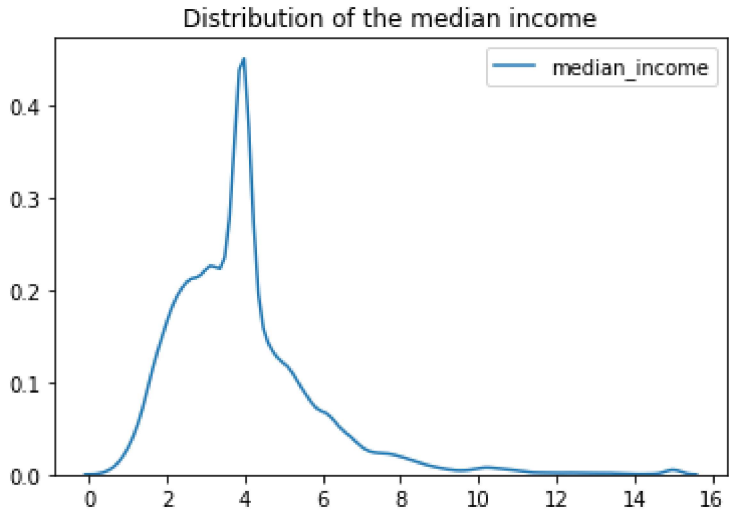


In [109]:
```
plt.subplot(2,1,1)
plt.title("Distribution of population")
sns.kdeplot(df.population)
plt.show()
plt.subplot(2,1,2)
plt.title("Distribution of Householders")
sns.kdeplot(df.households)
plt.show()
```





In [103]:
```
plt.title('Distribution of the median income')
sns.kdeplot(df.median_income)
plt.show()
```

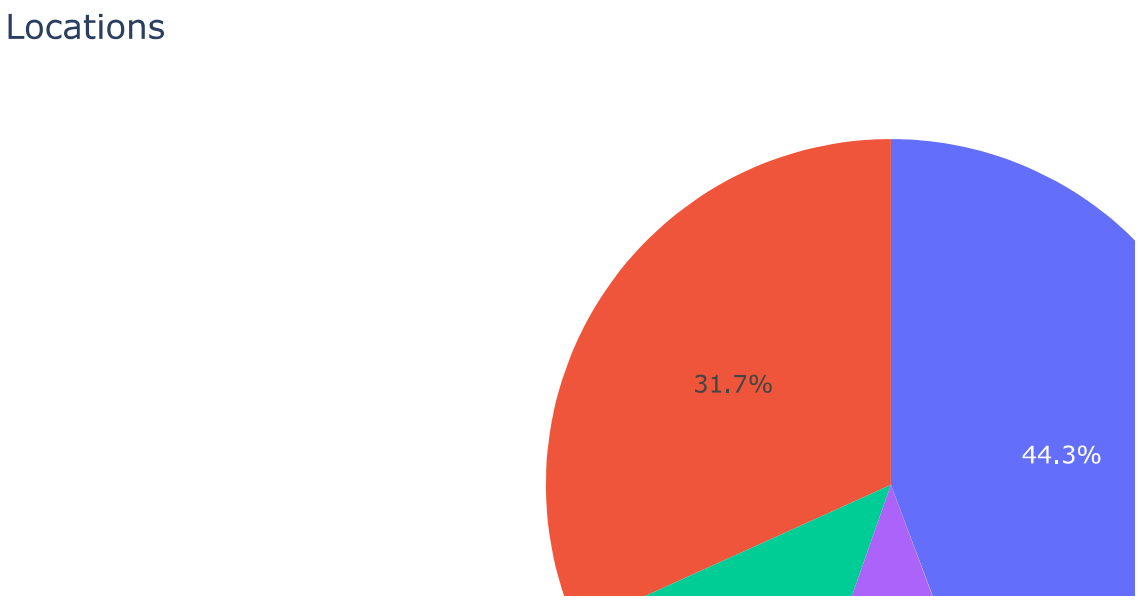


In [104]:
```
df.ocean_proximity.value_counts()
```

Out[104]:
```
<1H OCEAN     9136
INLAND        6551
NEAR OCEAN    2658
NEAR BAY      2290
ISLAND           5
Name: ocean_proximity, dtype: int64
```

In [106]:
```
import plotly.express as ex
```

In [107]:
```
ex.pie(df,names='ocean_proximity',title =' Locations')
```



In [ ]:
```

```