import matplotlib.pyplot as pyplot

#pyplot.plot(y)

#pyplot.ylabel("y")

#pyplot.show()

x = [1, 2, 3, 4, 5]

y = [1, 2, 3, 4, 5]

a = [1, 4, 9, 16, 25]

pyplot.plot(x, y, 'rs', x, a, 'b^')

pyplot.ylabel("y")

pyplot.xlabel("x")

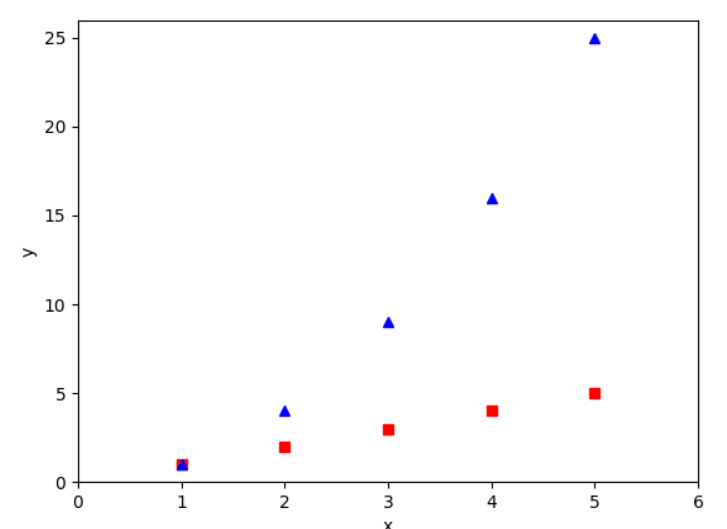
pyplot.axis([0, 6, 0, 26])

pyplot.show()

Problem

Read the data from the weight-height.csv file (use pandas read\_csv).

Plot weight as a function of height each dot being a blue circle.



|  |  |
| --- | --- |
| Group | Picture |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |

Code:

data = pd.read\_csv('weight-height.csv')

y = data['Height']

x = data['Weight']

pyplot.plot(x, y, 'bo')

pyplot.ylabel("Height")

pyplot.xlabel("Weight")

pyplot.show()

**Problem 2**

Calculate a linear model of Weight v Height and plot the line as a red straight line.

|  |  |
| --- | --- |
| Group | Picture |
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |

Group 2 code:

import matplotlib.pyplot as pyplot

import pandas as pd

import numpy as np

data = pd.read\_csv("C:\\data analysis\\weight-height.csv")

model = np.polyfit(data["Height"], data["Weight"], 1)

predict = np.poly1d(model)

x\_lin\_reg = range(54, 80)

y\_lin\_reg = predict(x\_lin\_reg)

pyplot.plot(data["Height"], data["Weight"], "bo", x\_lin\_reg, y\_lin\_reg, "r")

pyplot.xlabel("Height")

pyplot.ylabel("Weight")

pyplot.show()

import matplotlib.pyplot as pyplot

import pandas as pd

import statsmodels.api as sm

data = pd.read\_csv("weight-height.csv")

# pyplot.plot(data["Height"], data["Weight"], "bo")

# pyplot.xlabel("Height")

# pyplot.ylabel("Weight")

# pyplot.show()

print(data)

x = data["Height"]

Y = data["Weight"]

gender = data["Gender "]

numeric\_gender = gender == "Male"

X = sm.add\_constant(x)

model = sm.OLS(Y,X).fit()

prediction = model.predict()

pyplot.plot(x, Y, "bo", x, prediction, "r")

pyplot.show()

pyplot.scatter(x, Y, c=numeric\_gender, alpha=0.3)

pyplot.show()

pyplot.hist(x)

pyplot.show()1

Break 10 min - Back at 7:30

If x is continuous → Scatter Plot

If x is discrete → Box Plot

Problem 3

Using the correct plot. *Identify* the effect of Zip Code (pick 3 zip codes), sqft\_living, and #bedrooms.

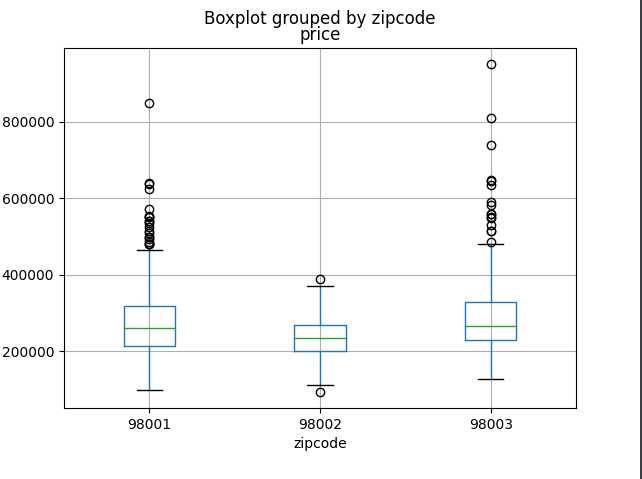
Run a regression of price vs all of them, see if the p-values align with your conclusion from the chart.

For the **sqft\_living** plot the prediction line

**Group 1**

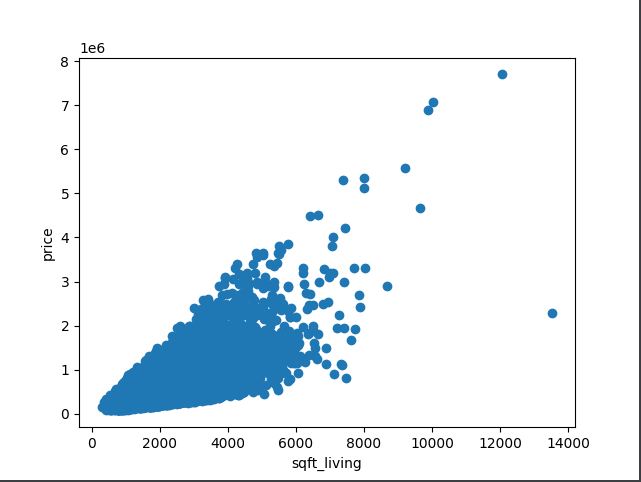
Plot of price v Zip (your choice of 3)

[Plot1]



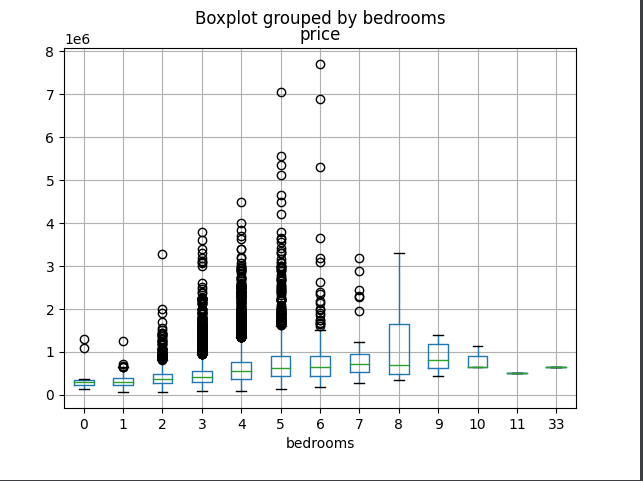
Plot of price v sqft\_living

[Plot2]



Plot of price v bedrooms

[Plot3]



Plot of prive v sqft\_living with prediction line

[Plot4]

**Group 2**

Plot of price v Zip (your choice of 3)

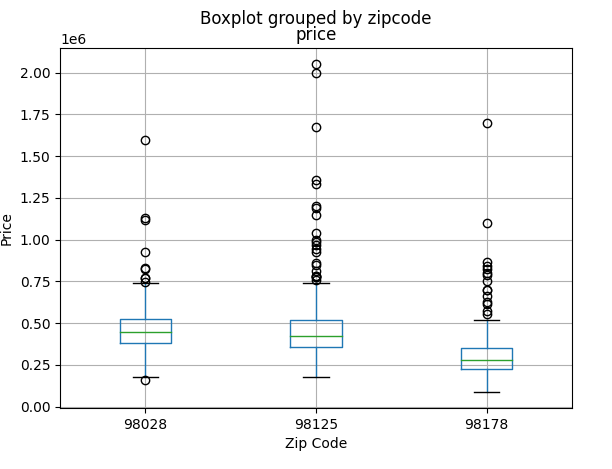
zipdata = data[(data['zipcode'] == 98178) | (data['zipcode'] == 98125) | (data['zipcode'] == 98028)]

zipdata.boxplot(column = "price", by = "zipcode")

pyplot.xlabel("Zip Code")

pyplot.ylabel("Price")

pyplot.show()



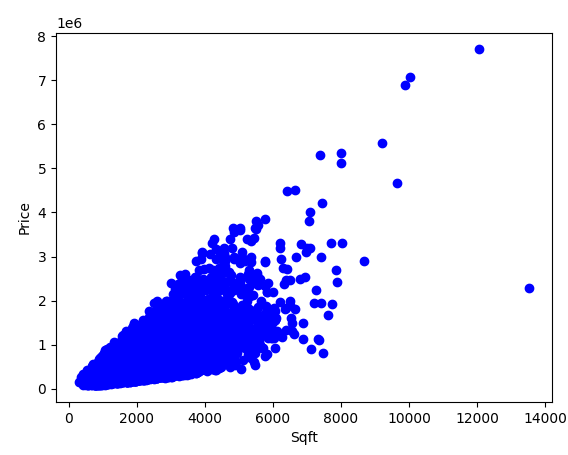
Plot of price v sqft\_living

pyplot.plot(data["sqft\_living"], data["price"], "bo")

pyplot.xlabel("Sqft")

pyplot.ylabel("Price")

pyplot.show()



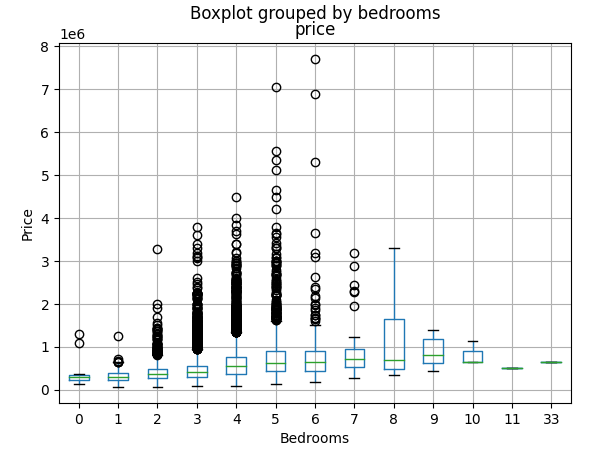
Plot of price v bedrooms

data.boxplot(column = "price", by = "bedrooms")

pyplot.xlabel("Bedrooms")

pyplot.ylabel("Price")

pyplot.show()



Plot of prive v sqft\_living with prediction line

model = np.polyfit(data["sqft\_living"], data["price"], 1)

predict = np.poly1d(model)

x\_lin\_reg = range(min(data["sqft\_living"]), max(data["sqft\_living"]))

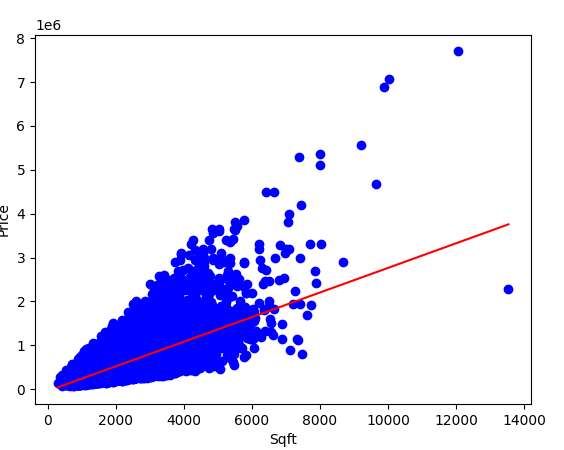
y\_lin\_reg = predict(x\_lin\_reg)

pyplot.plot(data["sqft\_living"], data["price"], "bo", x\_lin\_reg, y\_lin\_reg, "r")

pyplot.xlabel("Sqft")

pyplot.ylabel("Price")

pyplot.show()



**Group 3**

Plot of price v Zip (your choice of 3)

[Plot1]

Plot of price v sqft\_living

[Plot2]

Plot of price v bedrooms

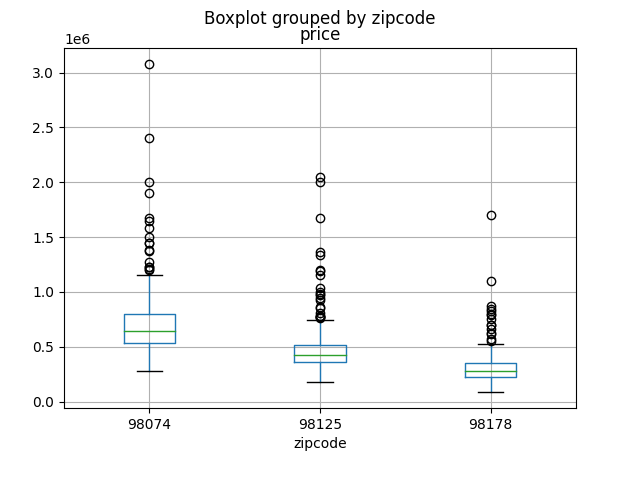
[Plot3]

Plot of prive v sqft\_living with prediction line

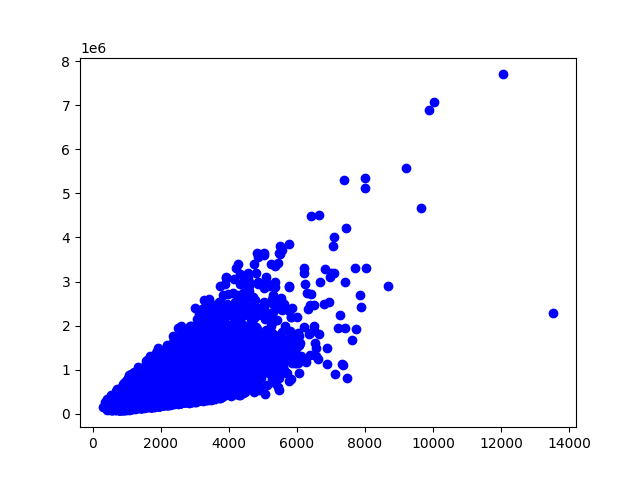
[Plot4]

**Group 4**

Plot of price v Zip (your choice of 3)



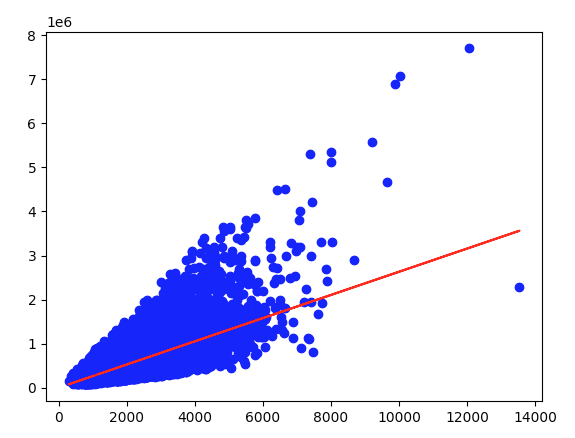
Plot of price v sqft\_living



Plot of price v bedrooms

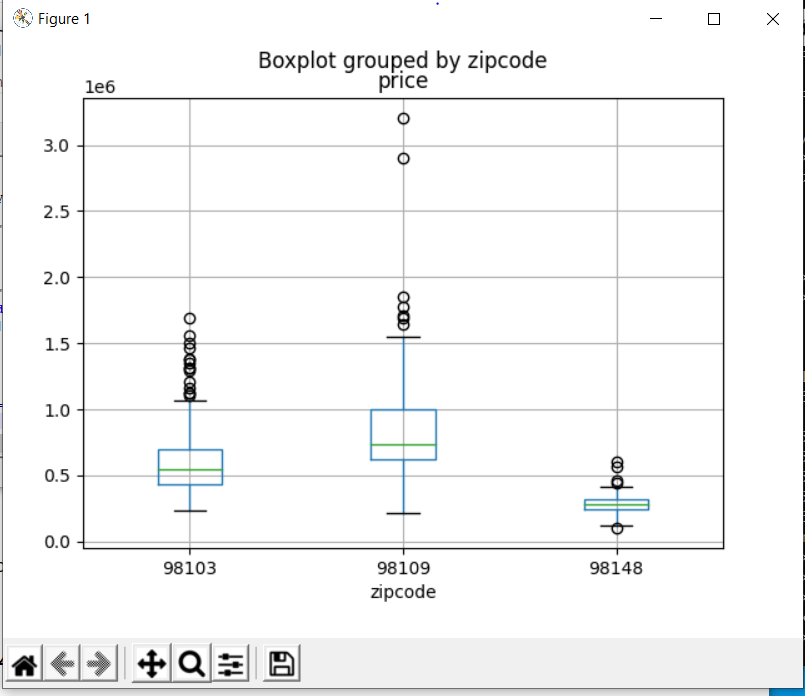
[Plot3]

Plot of prive v sqft\_living with prediction line



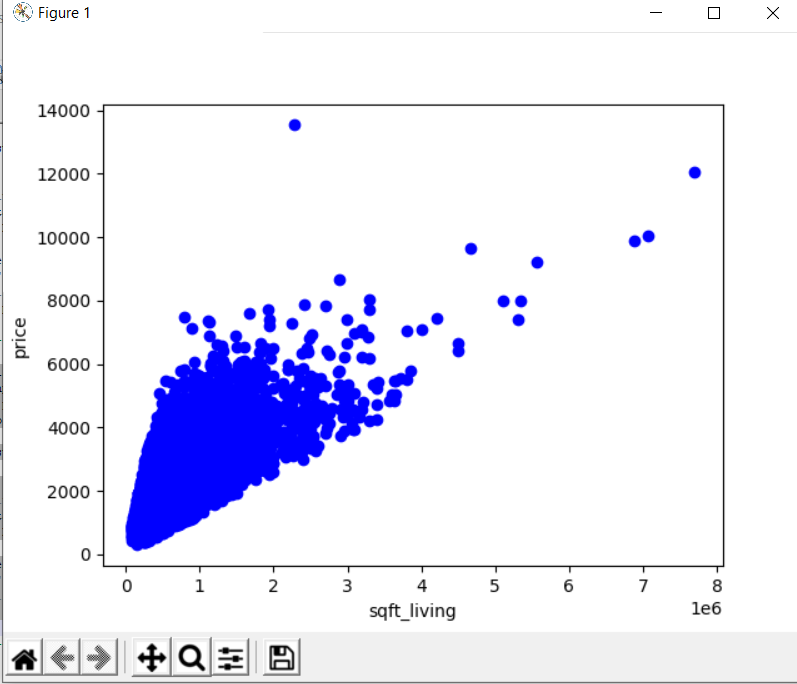
**Group 5**

Plot of price v Zip (your choice of 3)



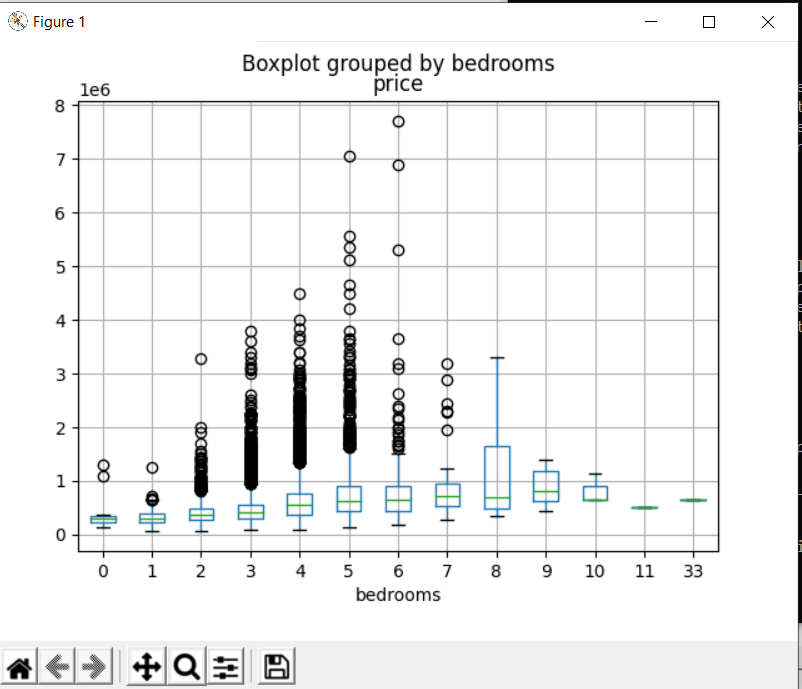
[Plot1]

Plot of price v sqft\_living



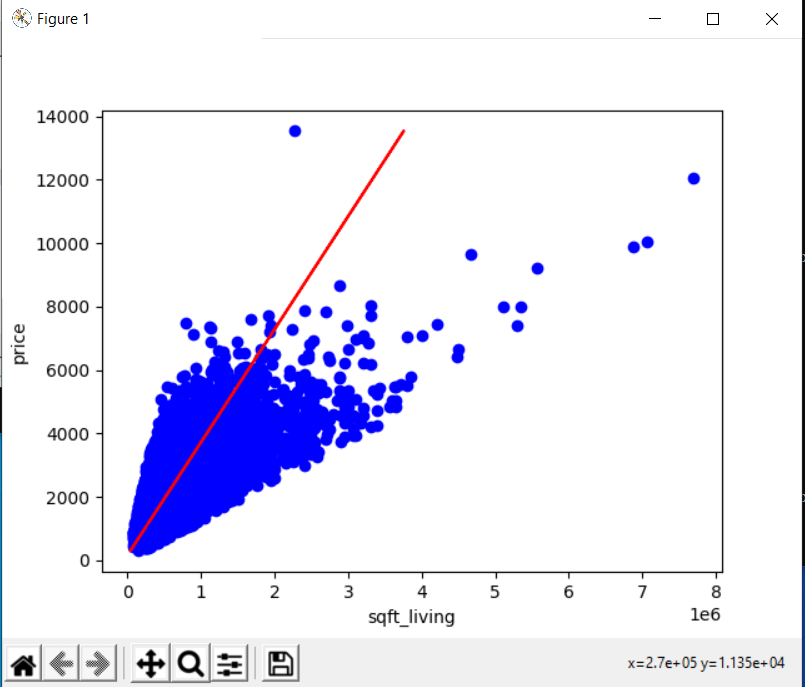
[Plot2]

Plot of price v bedrooms



[Plot3]

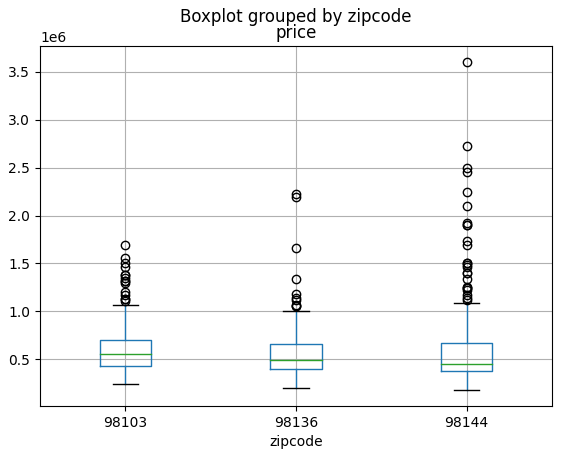
Plot of prive v sqft\_living with prediction line



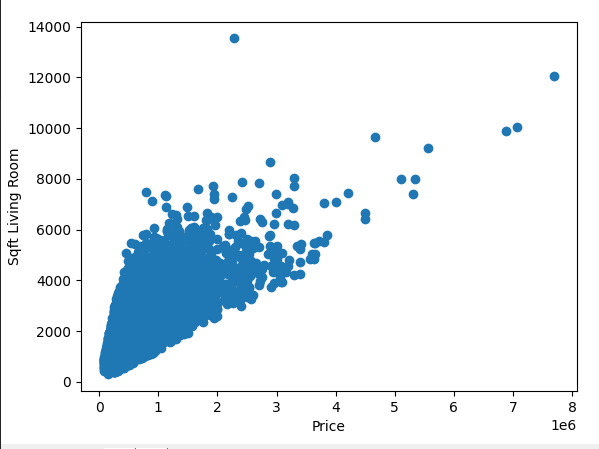
[Plot4]

**Group 6**

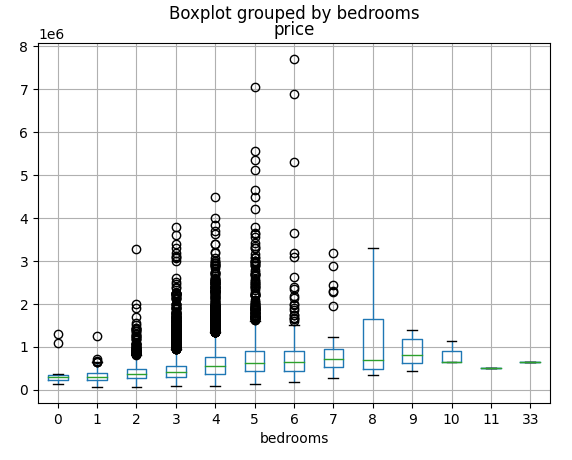
Plot of price v Zip (your choice of 3)

[Plot1] 

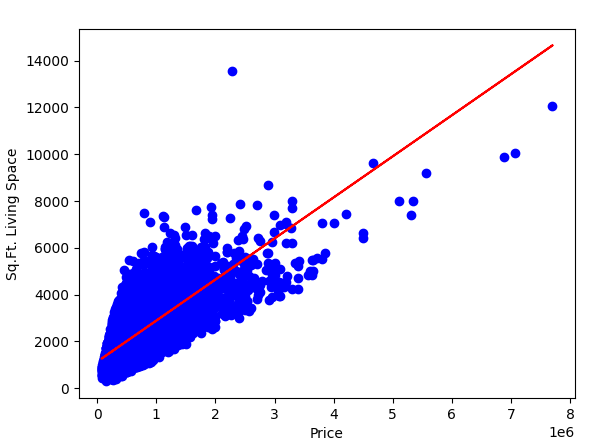
Plot of price v sqft\_living



Plot of price v bedrooms

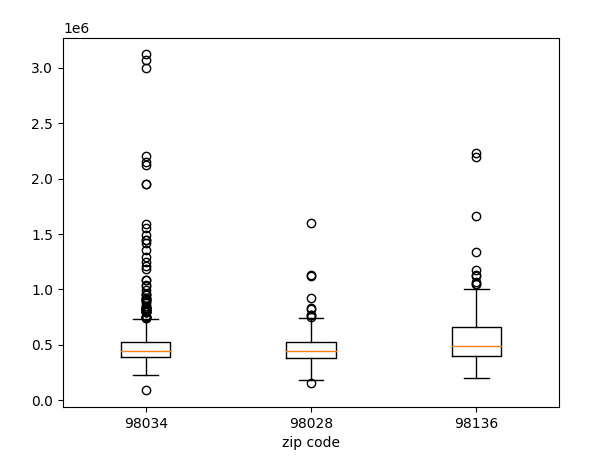
[Plot3]

Plot of prive v sqft\_living with prediction line

[Plot4]

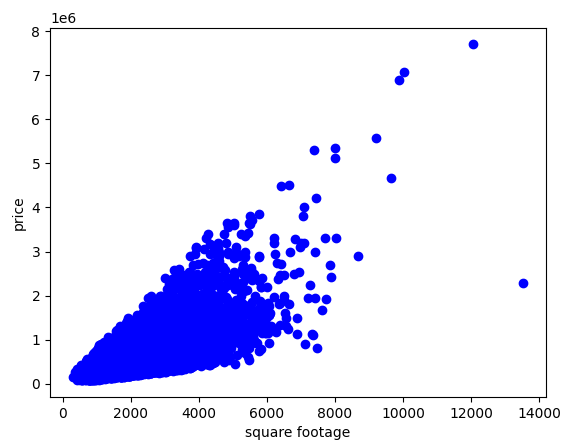
**Group 7**

Plot of price v Zip (your choice of 3)



[Plot1]

Plot of price v sqft\_living

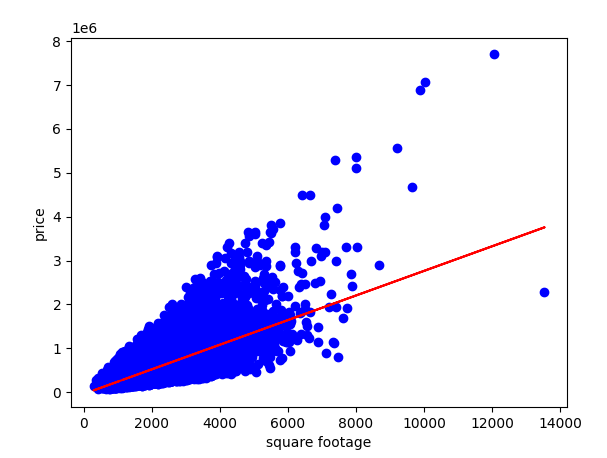


[Plot2]

Plot of price v bedrooms

[Plot3]

Plot of prive v sqft\_living with prediction line



[Plot4]

**Final Projects**

**Final Project (Part 1) - A series of data analysis questions in lestusedata, this will be based on googleplaystore that is in the same database as we used before.**

**Final Project (Part 2) - A short PowerPoint that explains the results of Part 1 (including charts) will be in Canvas.**

**[Hw8 / Hw9 / Final] - 1.5 weeks to finish from 3/14 till 3/22 for the code and 3/23 for presentation**

**Next Week 3/16 - Synchronous class (normal)**

**Week After 3/23 - No synchronous class (spend the time finishing your final)**

**Final due date is FINAL. Grades are due on 3/26.**