14.1 HW1_LAB1: Creating correlation matrices

The kc*house*data dataset contains information on house sale prices in King County, Washington from May 2014 and May 2015. The columns include sale price, number of bedrooms, and square footage of living space.

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

```
Load the data set into a data frame.
Find the correlation matrix for the price, bedrooms, and sqft_living
columns, in that order.
```

Ex: If the number of bathrooms and square footage of the lot, bathrooms and sqft*lot, are used instead of bedrooms and sqft*living, the output is:

```
                price   bathrooms  sqft_lot
price        1.000000    0.525138  0.089661
bathrooms    0.525138    1.000000  0.087740
sqft_lot     0.089661    0.087740  1.000000
```

NaN.2635824.qx3zqy7

LAB ACTIVITY | 14.1.1: HW1_LAB1: Creating correlation matrices | 5 / 5

Downloadable files

`kc_house_data.csv` **Download**

main.py — Load default template...

```
1  # Import the necessary modules
2  import pandas as pd
3
4  # Code to read in kc_house_data.csv"
5  house = pd.read_csv('kc_house_data.csv')
6
7  # Display the correlation matrix for the columns price, bedrooms, and sqft_living, in
8  # Code to calculate correlation matrix
9  print(house[['price', 'bedrooms', 'sqft_living']].corr())
10
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Develop mode | Submit mode

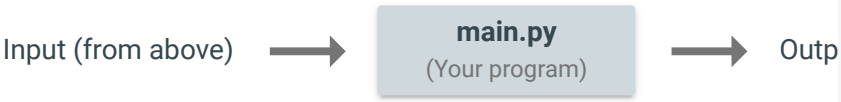
Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Run program



Program output displayed here

Coding trail of your work What is this?

```
6/12 U----------------0,5 min:23
```

14.2 HW1_LAB2: Creating SLR models using linregress()

The kc*house*data dataset contains information on house sale prices in King County, Washington from May 2014 and May 2015. The columns include sale price, number of bedrooms, and square footage of living space.

```
Load the data set into a data frame.
Find the correlation matrix for the price, bedrooms, and sqft_living
columns.
Set y = house.price and x equal to the column with the highest
correlation.
Use the linregress() function to perform a simple linear regression on
x and y.
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Ex: If the number of bathrooms and square footage of the lot, bathrooms and sqft*lot, are used instead of bedrooms and sqft*living, the output is:

```
             price  bathrooms  sqft_lot
price      1.000000   0.525138  0.089661
bathrooms  0.525138   1.000000  0.087740
sqft_lot   0.089661   0.087740  1.000000
LinregressResult(slope=280.6235678974481, intercept=-43580.74309447361,
rvalue=0.7020350546117996, pvalue=0.0, stderr=1.9363985519989153,
intercept_stderr=4402.689690303904)
```

©zyBooks 06/14/22 19:37 1317912 Lee Brunovsky GSWCSCI6420MukkavilliSummer2022

NaN.2635824.qx3zqy7

LAB ACTIVITY | 14.2.1: HW1_LAB2: Creating SLR models using linregress() 5 / 5

Downloadable files

`kc_house_data.csv` **Download**

main.py Load default template...

```
# Import the necessary modules
import pandas as pd
import scipy.stats as st

# Code to read in kc_house_data.csv"
house = pd.read_csv('kc_house_data.csv')

# Display the correlation matrix for the columns price, bedrooms, and sqft_living, in
# Code to calculate correlation matrix
print(house[['price', 'bedrooms', 'sqft_living']].corr())

# Perform simple linear regression on price and the column with the highest correlati
# Code to perform SLR using linregress
model = st.linregress(house['sqft_living'], house['price'])

# Print the summary
print(model)
```

©zyBooks 06/14/22 19:37 1317912 Lee Brunovsky GSWCSCI6420MukkavilliSummer2022

Develop mode | **Submit mode**

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

Run program Input (from above) → main.py (Your program) → Outp

Program output displayed here

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Coding trail of your work What is this?

```
6/12 U-------0,0---5 min:15
```

14.3 HW1_LAB3: Making predictions using SLR models

The kc_house_data dataset contains information on house sale prices in King County, Washington from May 2014 and May 2015. The columns include sale price, number of bedrooms, and square footage of living space.

Write a program using the ols() function that creating a model that takes in the square footage of living space in a house and gives the price of the house as output.

For example, if the input is:

```
1000
```

the output should be:

```
0    237042.824803
dtype: float64
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

NaN.2635824.qx3zqy7

LAB ACTIVITY 14.3.1: HW1_LAB3: Making predictions using SLR models 15 / 15

Downloadable files

`kc_house_data.csv` **Download**

main.py Load default template...

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as sms

# load the file kc_house_data.csv
house = pd.read_csv("kc_house_data.csv")

# fit a linear model using the sms.ols function and the house dataframe
model = sms.ols('price ~ sqft_living', house).fit()

area = float(input())

# use the model.predict function to find the predicted value for price using
## the area value for the predictor
prediction = model.predict(pd.DataFrame(np.array([[area]]),columns=['sqft_living']))

print(prediction)
```

©zyBooks 06/14/22 19:37 1317912 Lee Brunovsky GSWCSCI6420MukkavilliSummer2022

Develop mode | **Submit mode**

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

1000

Run program Input (from above) → **main.py** (Your program) → Outp

Program output displayed here

Coding trail of your work What is this?

```
6/12 U-----------------------------------------------------
---------0,15 min:81
```

©zyBooks 06/14/22 19:37 1317912 Lee Brunovsky GSWCSCI6420MukkavilliSummer2022

14.4 HW1_LAB4: Multiple regression

The kc_house_data dataset contains information on house sale prices in King County, Washington from May 2014 and May 2015. The columns include sale price, square footage of living space, and condition of the house.

- Load the data set into a data frame.
- Use the ols function to perform a multiple linear regression with price as the response variable and sqft_living and condition as the predictor variables.
- Create an analysis of variance table using the results of the multiple regression.

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Ex: If bedrooms is used instead of condition the output is:

```
                   sum_sq       df             F         PR(>F)
sqft_living  1.199317e+15      1.0  18040.167832   0.000000e+00
bedrooms     4.063538e+13      1.0    611.238731   4.244531e-133
Residual     1.436641e+15  21610.0           NaN            NaN
```

NaN.2635824.qx3zqy7

LAB ACTIVITY 14.4.1: HW1_LAB4: Multiple regression 5 / 5

Downloadable files

kc_house_data.csv **Download**

main.py Load default template...

```
1  # Import the necessary modules
2  import pandas as pd
3  import statsmodels.api as sm
4  from statsmodels.stats.anova import anova_lm
5  from statsmodels.formula.api import ols
6
7  # Code to read in kc_house_data.csv
8  housing = pd.read_csv('kc_house_data.csv')
9
10 results = ols('price ~ sqft_living +(condition)', housing).fit()
11
12 # Create an analysis of variance table
13 # Code to create ANOVA table
14 aov_table = sm.stats.anova_lm(results, typ=2)
15
16 # Print the analysis of variance table
17 print(aov_table)
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Develop mode Submit mode

Develop mode | Submit mode

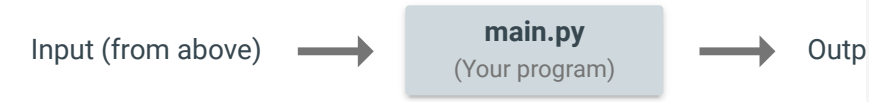
Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Run program



Program output displayed here

Coding trail of your work What is this?

```
6/13 M----------------------------------------0-5 min:49
```

14.5 HW1_LAB5: Making predictions using MLR models

The wine dataset contains 4,898 instances with 11 features that include pH, percentage of alcohol, and others, used to predict wine quality. Constructing a box plot for the free sulphur dioxide feature shows a right-skewed distribution.

Write a program that creates a model that uses chlorides, pH, and alcohol as predictor variables and returns the quality as the response variable.

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

If the inputs are:

```
0.03
4
10
```

the output should be:

```
0    5.975599
dtype: float64
```

NaN.2635824.qx3zqy7

LAB ACTIVITY | 14.5.1: HW1_LAB5: Making predictions using MLR models | 15 / 15

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Downloadable files

wine.csv **Download**

main.py Load default template...

```
1  # load the necessary modules
2  import pandas as pd
3  import statsmodels.api as sm
4  import numpy as np
5  from statsmodels.formula.api import ols
6
7  # load wine.csv
8  wine = pd.read_csv('wine.csv')
9
10 x = wine[['chlorides', 'pH','alcohol']]
11 y = wine['quality']
12
13 # fit a multiple regression model using the sms.ols function
14 model = ols('quality ~ chlorides+pH+alcohol', wine).fit()
15
16 chlorides_in = float(input())
17 pH_in = float(input())
18 alcohol_in = float(input())
```

Develop mode | **Submit mode**

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

```
0.03
4
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Run program Input (from above) → **main.py** (Your program) → Outp

Program output displayed here

Coding trail of your work What is this?

```
6/13 M-------------------------------------------------------
-----0-15 min:72
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

14.6 HW1_LAB6: Calculating VIF using variance_inflation_factor()

The kc_house_data dataset contains information on house sale prices in King County, Washington from May 2014 and May 2015. The columns include sale price, and a number of variables that might affect the price.

- Load the data set into a data frame.
- Create matrices y, from the price column, and X, from the bedrooms, bathrooms, sqft_living, sqft_lot, yr_built, and sqft_living15 columns.
- calculate the VIF for each predictor variable.

Ex: If sqft_lot15 is used instead of sqft_living15 the output is:

```
[1.563916, 2.961241, 2.742922, 2.076869, 1.382412, 2.094995]
```

NaN.2635824.qx3zqy7

LAB ACTIVITY 14.6.1: HW1_LAB6: Calculating VIF using variance_inflation_factor() 5 / 5

Downloadable files

kc_house_data.csv **Download**

main.py Load default template...

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

```
1 # import the necessary modules
2 import pandas as pd
3 from statsmodels.stats.outliers_influence import variance_inflation_factor
4 from patsy import dmatrices
5
6 # read in the csv file
7 house = pd.read_csv('kc_house_data.csv')
8
9 # create matrices y, from the price column, and X, from the bedrooms, bathrooms, saft
```

```
y, X = dmatrices('price ~ bedrooms+bathrooms+sqft_living+sqft_lot+yr_built+sqft_livin
# calculate the VIF for each X_i
vif = [variance_inflation_factor(X.values, i+1) for i in range(X.shape[1]-1)]
result = [round(num, 6) for num in vif]
print(result)
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky

Develop mode | Submit mode

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

Run program

Input (from above) → **main.py** (Your program) → Outp

Program output displayed here

Coding trail of your work What is this?

6/14 T---5 min:7

14.7 HW1_LAB7: Confidence and prediction intervals for MLR models

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

The wine dataset contains 4,898 instances with 11 features that include pH, percentage of alcohol, and others, used to predict wine quality. Constructing a box plot for the free sulphur dioxide feature shows a right-skewed distribution.

Write a program that outputs the confidence and prediction intervals of a model that uses chlorides, pH, and alcohol as predictor variables and returns the quality as the response variable, given inputs for

the predictor variables.

If the inputs are:

```
0.03
4
10
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

the output should be:

```
mean    mean_se   mean_ci_lower   mean_ci_upper   obs_ci_lower
obs_ci_upper
0   5.975599   0.063751  5.850619   6.10058   4.412228   7.53897
```

NaN.2635824.qx3zqy7

LAB ACTIVITY | 14.7.1: HW1_LAB7: Confidence and prediction intervals for MLR models 15 / 15

Downloadable files

wine.csv **Download**

main.py Load default template...

```
# load the necessary modules
import pandas as pd
import statsmodels.formula.api as sm
from statsmodels.formula.api import ols

# load wine.csv
wine = pd.read_csv('wine.csv')

# fit a multiple regression model using the sms.ols function
model = ols('quality ~ chlorides+pH+alcohol', wine).fit()


chlorides = float(input())
pH = float(input())
alcohol = float(input())

# create a pandas dataframe with chlorides, pH, and alcohol as columns
# with 1 row representing an instance with the given inputs for each column
```

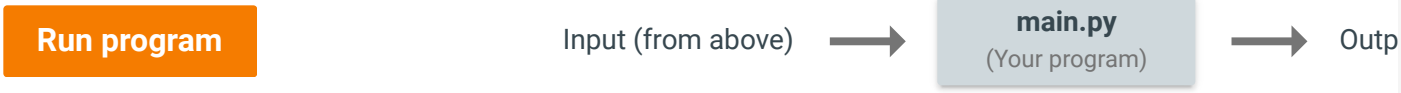
©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Develop mode **Submit mode**

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

0.03
4



Program output displayed here

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022

Coding trail of your work What is this?

```
6/14 T------15 min:7
```

©zyBooks 06/14/22 19:37 1317912
Lee Brunovsky
GSWCSCI6420MukkavilliSummer2022