Kernel: Python 3 (system-wide)

In [1]:
```
# computational imports
import statsmodels.api as sm
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
# for reading files from urls
import urllib.request
# display imports
from IPython.core.display import HTML

# import notebook styling for tables and width etc.
response =
urllib.request.urlopen('https://raw.githubusercontent.com/DataScienceU
WL/DS775v2/master/ds755.css')
HTML(response.read().decode("utf-8"));
```

Project 1: Report

Use this Jupyter notebook to summarize the details of this project organized in the following sections. Note, there is also a presentation notebook that accompanies this project.

The file `Airfares.xlsx` contains real data that were collected between Q3-1996 and Q2-1997. The first sheet contains variable descriptions while the second sheet contains the data. A csv file of the data is also provided (called *Airfares.csv*).

To get full credit your code should all run and produce correct answers if the data in the file `Airfares.xlsx` is changed. That means you can't type in coefficients for your linear models, but will have to store them in variables instead.

P1.1 - Introduction

Summarize the problem statement, establishing the context and methods used in this project. (Write an introduction that says what you're going to do and how you're going to do it!)

*** 5 points - answer in cell below *** (don't delete this cell)

The goal of this project is to maximize the average fares using three variables: coupons, Herfindel index, and distance of the route. This is to investigate ways to decrease airport congestion by possibly repurposing old military bases of smaller municipal airports. Linear regression will be used to determine relationships between the number of passengers on the route, the average income of the departure city, and the average income of the arrival city. A model will also be created to determine the relationship between the average fare, coupons, the Herfindel Index and the distance of the route. These will be used to create a linear programming model to determine the optimal price point for the fare of the flight.

P1.2 - Linear Regression Models

Provide a brief summary of the linear regression models used to estimate coefficients that will be used in the linear programming problem. Explain why the multiple regression equations had to be fitted through the origin (consider the assumptions of linear programming).

*** 5 points - answer in cell below *** (don't delete this cell)

In [2]:
```
# code for linear regression models goes here
air_fares = pd.read_csv("data/Airfares.csv")

# define predictor variables
X_FARE = air_fares[['COUPON', 'HI', 'DISTANCE']]

# define response variables
Y_FARE = air_fares['FARE']

# Fit the objective function and pull out coefficients
model_FARE = sm.OLS(Y_FARE,X_FARE).fit()
coefs_FARE = model_FARE.params

print(model_FARE.summary())
print(coefs_FARE)
```

Out[2]:

```
                                 OLS Regression Results
=========================================================================
===============
Dep. Variable:                   FARE   R-squared (uncentered):
0.911
Model:                            OLS   Adj. R-squared (uncentered):
0.911
Method:                 Least Squares   F-statistic:
2165.
Date:                Sun, 04 Oct 2020   Prob (F-statistic):
0.00
Time:                        04:48:36   Log-Likelihood:
-3439.5
No. Observations:                 638   AIC:
6885.
Df Residuals:                     635   BIC:
6898.
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================
======
                 coef    std err          t      P>|t|      [0.025
0.975]
-------------------------------------------------------------------------
------
COUPON        22.5900      6.697      3.373      0.001       9.440
35.740
HI             0.0118      0.001     10.599      0.000       0.010
0.014
DISTANCE       0.0833      0.004     18.991      0.000       0.075
0.092
=========================================================================
======
Omnibus:                       31.675   Durbin-Watson:
0.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
16.008
Skew:                           0.193   Prob(JB):
0.000334
Kurtosis:                       2.327   Cond. No.
1.54e+04
=========================================================================
======

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 1.54e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
COUPON      22.590019
HI           0.011798
DISTANCE     0.083336
dtype: float64
```

In [3]:

```
# regression for PAX
X_PAX = air_fares[['COUPON', 'HI', 'DISTANCE']]
```

```
Y_PAX = air_fares['PAX']


model_PAX = sm.OLS(Y_PAX,X_PAX).fit()
coefs_PAX = model_PAX.params

print(model_PAX.summary())
print(coefs_PAX)
```

Out[3]:

```
                            OLS Regression Results
=========================================================================
===============
Dep. Variable:                    PAX   R-squared (uncentered):
0.424
Model:                            OLS   Adj. R-squared (uncentered):
0.421
Method:                 Least Squares   F-statistic:
155.6
Date:                Sun, 04 Oct 2020   Prob (F-statistic):
1.32e-75
Time:                        04:48:40   Log-Likelihood:
-6993.6
No. Observations:                 638   AIC:
1.399e+04
Df Residuals:                     635   BIC:
1.401e+04
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================
======
                 coef    std err          t      P>|t|      [0.025
0.975]
-------------------------------------------------------------------------
------
COUPON      1.082e+04   1758.617      6.152      0.000    7365.921
1.43e+04
HI             0.2482      0.292      0.849      0.396      -0.326
0.822
DISTANCE      -2.2980      1.152     -1.994      0.047      -4.561
-0.035
=========================================================================
======
Omnibus:                      345.744   Durbin-Watson:
0.689
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
1848.009
Skew:                           2.508   Prob(JB):
0.00
Kurtosis:                       9.660   Cond. No.
1.54e+04
=========================================================================
======

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 1.54e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
COUPON      10819.328522
HI              0.248183
DISTANCE       -2.298017
dtype: float64
```

In [4]:

```
# regression for S_INCOME
X_S_INCOME = air_fares[['COUPON', 'HI', 'DISTANCE']]
Y_S_INCOME = air_fares['S_INCOME']
```

```
model_S_INCOME = sm.OLS(Y_S_INCOME,X_S_INCOME).fit()
coefs_S_INCOME = model_S_INCOME.params

print(model_S_INCOME.summary())
print(coefs_S_INCOME)
```

Out[4]:

```
                            OLS Regression Results
=========================================================================
===============
Dep. Variable:               S_INCOME   R-squared (uncentered):
0.966
Model:                            OLS   Adj. R-squared (uncentered):
0.966
Method:                 Least Squares   F-statistic:
6023.
Date:                Sun, 04 Oct 2020   Prob (F-statistic):
0.00
Time:                        04:48:42   Log-Likelihood:
-6359.1
No. Observations:                 638   AIC:
1.272e+04
Df Residuals:                     635   BIC:
1.274e+04
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================
======
                 coef    std err          t      P>|t|      [0.025
0.975]
-------------------------------------------------------------------------
------
COUPON      2.091e+04    650.471     32.145      0.000    1.96e+04
2.22e+04
HI             1.1146      0.108     10.309      0.000       0.902
1.327
DISTANCE      -2.8310      0.426     -6.642      0.000      -3.668
-1.994
=========================================================================
======
Omnibus:                        6.012   Durbin-Watson:
1.164
Prob(Omnibus):                  0.049   Jarque-Bera (JB):
6.730
Skew:                          -0.141   Prob(JB):
0.0346
Kurtosis:                       3.417   Cond. No.
1.54e+04
=========================================================================
======

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 1.54e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
COUPON      20909.191409
HI              1.114583
DISTANCE       -2.830983
dtype: float64
```

In [5]:

```
# regression for E_INCOME
X_E_INCOME = air_fares[['COUPON', 'HI', 'DISTANCE']]
Y_E_INCOME = air_fares['E_INCOME']
```

```
model_E_INCOME = sm.OLS(Y_E_INCOME, X_E_INCOME).fit()
coefs_E_INCOME = model_E_INCOME.params

print(model_E_INCOME.summary())
print(coefs_E_INCOME)
```

Out[5]:

```
                            OLS Regression Results
==============================================================================
===============
Dep. Variable:               E_INCOME   R-squared (uncentered):
0.962
Model:                            OLS   Adj. R-squared (uncentered):
0.961
Method:                 Least Squares   F-statistic:
5288.
Date:                Sun, 04 Oct 2020   Prob (F-statistic):
0.00
Time:                        04:48:45   Log-Likelihood:
-6400.3
No. Observations:                 638   AIC:
1.281e+04
Df Residuals:                     635   BIC:
1.282e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
======
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
------
COUPON      1.833e+04    693.900     26.416      0.000     1.7e+04
1.97e+04
HI             1.4069      0.115     12.198      0.000       1.180
1.633
DISTANCE      -1.0198      0.455     -2.243      0.025      -1.913
-0.127
==============================================================================
======
Omnibus:                        4.753   Durbin-Watson:
0.540
Prob(Omnibus):                  0.093   Jarque-Bera (JB):
4.842
Skew:                           0.207   Prob(JB):
0.0888
Kurtosis:                       2.898   Cond. No.
1.54e+04
==============================================================================
======

Notes:
[1] R² is computed without centering (uncentered) since the model does
not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[3] The condition number is large, 1.54e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
COUPON      18330.370962
HI              1.406882
DISTANCE       -1.019802
dtype: float64
```

*** 5 points - answer in cell below *** (don't delete this cell)

Regression results for FARE: ParseError: KaTeX parse error: Undefined control sequence: \$ at position 59: …833 * DISTANCE \$ Regression re…PAX = 1.082 x 10^4 * COUPON + 0.2482 * HI - 2.2980 * DISTANCE\$ Regression results for S_INCOME: ParseError: KaTeX parse error: Undefined control sequence: \$ at position 68: …8310 * DISTANCE\$ Regression res…E_INCOME = 1.833 x 10^4 * COUPON + 1.4069 * HI - 1.0198 * DISTANCE\$ The R-squared values for FARE, S_INCOME, and E_INCOME are all close to 1, showing that these models have a good fit. The R-squared value for PAX, however, is less than 0.5 showing that we could do a better job with this model in the future.

The standard error in all models is close to 0 for Herfindel Index and distance of route, showing that the coefficients are more than likely accurate. The standard errors for coupon for all models are very high though, showing that there is a strong chance that these coefficients are not accurate.

The p-values for all predictor variables in all models are less than 0.05. This means that all over these predictors are statistically significant over all models.

The multiple regression equations had to be fitted through the feasible region because of the possibility that the response variables could reasonably be valued at 0

P1.3 - Optimal LP Solution

The optimal value of the airfare and for which values of COUPON, HI, and DISTANCE it occurs.

*** 8 points - answer in cell below *** (don't delete this cell)

In [15]:
```
# code for Pyomo and nicely formatted output goes here
from pyomo.environ import *

# Concrete Model
model = ConcreteModel(name = "Air Fare")

# Decision Variables
model.COUPON = Var(domain=NonNegativeReals, bounds=(0,1.5))
model.HI = Var(domain=NonNegativeReals, bounds=(4000,8000))
model.DISTANCE = Var(domain=NonNegativeReals, bounds=(500,1000))

# Objective
model.obj = Objective(expr=model.COUPON*coefs_FARE[0] +
model.HI*coefs_FARE[1]+model.DISTANCE*coefs_FARE[2], sense=maximize)

# Constraints
model.PAX_con = Constraint(expr=model.COUPON*coefs_PAX[0] +
model.HI*coefs_PAX[1] + model.DISTANCE*coefs_PAX[2] <= 20000)
model.S_INCOME_con = Constraint(expr=model.COUPON*coefs_S_INCOME[0] +
model.HI*coefs_S_INCOME[1]+model.DISTANCE*coefs_S_INCOME[2] <= 30000)
model.E_INCOME_con = Constraint(expr=model.COUPON*coefs_E_INCOME[0] +
model.HI*coefs_E_INCOME[1]+model.DISTANCE*coefs_E_INCOME[2] >= 30000)

# Solve
solver = SolverFactory('glpk')
```

```
solver.solve(model)

# display
print(f"Optimal Fare = ${model.obj():,.2f}")
```

Out[15]: Optimal Fare = $203.55

P1.4 - Sensitivity Report

From the sensitivity report, explain which constraints are binding for the number of passengers on that route (PAX), the starting city's average personal income (S_INCOME), and the ending city's average personal income (E_INCOME). If the constraint is binding, interpret the shadow price in the context of the problem. If the constraint is not binding, interpret the slack in the context of the problem.

*** 5 points - answer in cell below *** (don't delete this cell)

In [17]:
```
# code to generate and display sensitivity report goes here
# write the model to a sensitivity report
model.write('AirFares.lp', io_options={'symbolic_solver_labels':
True})
!glpsol -m AirFares.lp --lp --ranges sensitair.sen
f=open('sensitair.sen', 'r')
file_contents = f.read()
print(file_contents)
f.close()
```

Out[17]:
```
GLPSOL: GLPK LP/MIP Solver, v4.65
Parameter(s) specified in the command line:
 -m AirFares.lp --lp --ranges sensitair.sen
Reading problem data from 'AirFares.lp'...
4 rows, 4 columns, 10 non-zeros
36 lines were read
GLPK Simplex Optimizer, v4.65
4 rows, 4 columns, 10 non-zeros
Preprocessing...
2 rows, 3 columns, 6 non-zeros
Scaling...
 A: min|aij| =  1.020e+00  max|aij| =  2.091e+04  ratio =  2.050e+04
GM: min|aij| =  7.309e-01  max|aij| =  1.368e+00  ratio =  1.872e+00
EQ: min|aij| =  5.342e-01  max|aij| =  1.000e+00  ratio =  1.872e+00
Constructing initial basis...
Size of triangular part is 2
      0: obj =   8.885866366e+01 inf =   2.215e+04 (1)
      3: obj =   1.739717779e+02 inf =   0.000e+00 (0)
*     4: obj =   2.035540468e+02 inf =   0.000e+00 (0)
OPTIMAL LP SOLUTION FOUND
Time used:   0.0 secs
Memory used: 0.0 Mb (40412 bytes)
Write sensitivity analysis report to 'sensitair.sen'...
GLPK 4.65 - SENSITIVITY ANALYSIS REPORT
Page   1

Problem:
Objective:  obj = 203.5540468 (MAXimum)

   No. Row name     St      Activity         Slack   Lower bound
Activity      Obj coef  Obj value at Limiting
                                          Marginal   Upper bound
range         range   break point variable
------ ------------ -- ------------- ------------- -------------  ------
------- ------------- ------------- ------------
     1 c_u_PAX_con_ BS   12061.75912    7938.24088          -Inf
11353.40212       -.00209     178.36992 c_u_S_INCOME_con_
                                             .       20000.00000
12979.23503        .03224     592.46257 HI

     2 c_u_S_INCOME_con_
                    NU   30000.00000        .               -Inf
28631.04516       -.00108     202.07505 c_l_E_INCOME_con_
                                            .00108   30000.00000
37449.46803          +Inf     211.60236 COUPON

     3 c_l_E_INCOME_con_
                    BS   31200.11575   -1200.11575   30000.00000
10235.25243       -.00123     165.10359 c_u_S_INCOME_con_
                                             .              +Inf
31200.11575          +Inf          +Inf

     4 c_e_ONE_VAR_CONSTANT
                    NS       1.00000        .            1.00000
.                -Inf     203.55405 ONE_VAR_CONSTANT
                                             .           1.00000
+Inf          +Inf     203.55405

GLPK 4.65 - SENSITIVITY ANALYSIS REPORT
Page   2
```

```
Problem:
Objective:  obj = 203.5540468 (MAXimum)

   No. Column name  St      Activity      Obj coef   Lower bound
Activity      Obj coef  Obj value at Limiting
                                          Marginal   Upper bound
range         range   break point variable
------ ------------ -- ------------ ------------ ------------  ------
------- ------------ ------------ ------------
     1 COUPON       BS       1.14372      22.59002        .
1.07825        .          177.71733 c_u_S_INCOME_con_
                                            .            1.50000
1.29258     221.32046     430.84659 HI

     2 DISTANCE     NU    1000.00000        .08334     500.00000
179.14046       -.00306     132.63636 c_l_E_INCOME_con_
                                            .08639    1000.00000
3631.40702          +Inf      430.89292 COUPON

     3 HI           NU    8000.00000        .01180    4000.00000
5207.50816        .00120      173.97178 c_l_E_INCOME_con_
                                            .01059    8000.00000
29455.84475          +Inf      430.84659 COUPON

     4 ONE_VAR_CONSTANT
                    BS       1.00000        .             .
1.00000          -Inf          -Inf
                                            .               +Inf
1.00000          +Inf          +Inf
```

```
End of report
```

*** 5 points - answer in cell below *** (don't delete this cell)

In this model S_INCOME is a binding constraint. The fare would increase by \$0.00108 for every increase of in in S_INCOME. PAX and E_INCOME are non-binding constraints. PAX can be decreased by ~7948 passengers and E_INCOME can be increased by \$1200.12 without increaseing the cost of FARE

P1.5 - Allowable Ranges

Interpret the allowable ranges (objective coefficient range) for COUPON, HI, and DISTANCE in the context of the problem.

*** 5 points - answer in cell below *** (don't delete this cell)

The allowable range for PAX is between ~11353 and ~12979 passengers on the route
The allowable range for S_INCOME is between $28641.05 and $37449.47 for the departure city

The allowable range for E_INCOME is between $10235.25 and $31.200.12 for the arrival city

P1.6 - Conclusion

Briefly summarize the main conclusion of this project, state what you see as any limitations of the methods used here, and suggest other possible methods of addressing the maximizing of airfare in this problem scenario.

*** 7 points - answer in cell below *** (don't delete this cell)

From this analysis we can see that the optimal value for the fare when using coupons, Hefindel Index and distance of routes is \$203.55. The R-squared value for coupon was quite low and the standard error was quite high for coupon. This leads me to believe that we could improve this model more by selecting different variables or trying to improve our fit when using coupons. It would be interesting to see if the departure city and arrival cities population would play any role in maximizing our fare prices. Using the vacation variable to see if we could leverage some more value out of cities that are or are not considered vacation destinations.

P1.7 - Appendix

Show the mathematical formulation for the linear programming problem used in this project.

You can either use LaTeX and markdown or take a clean, cropped picture of neatly handwritten equations and drag-n-drop it here.

*** 5 points - answer in cell below *** (don't delete this cell)

Maximize: $FARE = 22.5900 * COUPON + 0.118 * HI + 0.0833 * DISTANCE\\$
Such that:

ParseError: KaTeX parse error: Undefined control sequence: \$ at position 74: …NCE \leq 20000 \$ S_INCOME = 2.091 x 10^4 * COUPON + 1.1146 * HI - 2.2980 * DISTANCE \leq 30000 $ ParseError: KaTeX parse error: Undefined control sequence: \$ at position 80: …NCE \geq 30000 \$ COUPON \leq 1.5 $ $4000 \leq HI \leq 8000 $ ParseError: KaTeX parse error: Undefined control sequence: \$ at position 29: …ANCE \leq 1000 \$ FARE, COUPON, E_INCOME, S_INCOME, HI, DISTANCE \geq 0$

In [0]:

```

```