CIS400Assignment2HuahaoShang

September 28, 2022

[1]:

```
# Generic inputs for most ML tasks
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.ensemble import RandomForestRegressor

pd.options.display.float_format = '{:,.2f}'.format

# setup interactive notebook mode
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

from IPython.display import display, HTML
```

[2]:

```
# fetch data

df = pd.read_csv('NZ_airfares.csv')

df.head()

print(df)
```

[2]:

```
  Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration  \
0  19/09/2019          AKL   1:35 PM          CHC   3:00 PM   1h 25m
1  19/09/2019          AKL   3:55 PM          CHC   5:20 PM   1h 25m
2  19/09/2019          AKL  11:40 AM          CHC   1:05 PM   1h 25m
3  19/09/2019          AKL   8:00 PM          CHC   9:25 PM   1h 25m
4  19/09/2019          AKL   9:00 AM          CHC  10:25 AM   1h 25m

     Direct Transit Baggage          Airline  Airfare(NZ$)
0  (Direct)     NaN     NaN          Jetstar           111
1  (Direct)     NaN     NaN          Jetstar           111
2  (Direct)     NaN     NaN          Jetstar           132
```

```
3  (Direct)     NaN     NaN          Jetstar          132
4  (Direct)     NaN     NaN  Air New Zealand          133

       Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration  \
0       19/09/2019          AKL   1:35 PM          CHC   3:00 PM   1h 25m
1       19/09/2019          AKL   3:55 PM          CHC   5:20 PM   1h 25m
2       19/09/2019          AKL  11:40 AM          CHC   1:05 PM   1h 25m
3       19/09/2019          AKL   8:00 PM          CHC   9:25 PM   1h 25m
4       19/09/2019          AKL   9:00 AM          CHC  10:25 AM   1h 25m
…            …          …       …          …       …      …
162828  18/12/2019          ZQN   4:55 PM          WLG  10:10 PM   5h 15m
162829  18/12/2019          ZQN   9:35 AM          WLG   3:10 PM   5h 35m
162830  18/12/2019          ZQN  10:20 AM          WLG   6:10 PM   7h 50m
162831  18/12/2019          ZQN  10:20 AM          WLG   6:40 PM   8h 20m
162832  18/12/2019          ZQN   9:35 AM          WLG   6:10 PM   8h 35m

          Direct        Transit Baggage          Airline  Airfare(NZ$)
0       (Direct)            NaN     NaN          Jetstar           111
1       (Direct)            NaN     NaN          Jetstar           111
2       (Direct)            NaN     NaN          Jetstar           132
3       (Direct)            NaN     NaN          Jetstar           132
4       (Direct)            NaN     NaN  Air New Zealand           133
…          …            …     …              …           …
162828  (1 stop)  2h 15m in AKL     NaN  Air New Zealand           422
162829  (1 stop)  2h 35m in AKL     NaN  Air New Zealand           422
162830  (1 stop)  4h 50m in AKL     NaN  Air New Zealand           422
162831  (1 stop)  5h 20m in AKL     NaN  Air New Zealand           422
162832  (1 stop)  5h 35m in AKL     NaN  Air New Zealand           422

[162833 rows x 11 columns]
```

[3]:

```
#Filt 'Air New Zealand' in Airline column. And then drop "airline"
df_sub = df[(df.Airline == 'Air New Zealand')]
df_sub = df_sub.drop('Airline',axis = 1)
df_sub.head()
print(df_sub)
```

[3]:

```
   Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration  \
4   19/09/2019          AKL   9:00 AM          CHC  10:25 AM   1h 25m
7   19/09/2019          AKL  11:00 AM          CHC  12:25 PM   1h 25m
8   19/09/2019          AKL   7:00 PM          CHC   8:25 PM   1h 25m
10  19/09/2019          AKL  10:00 AM          CHC  11:25 AM   1h 25m
11  19/09/2019          AKL   4:00 PM          CHC   5:25 PM   1h 25m

      Direct Transit Baggage  Airfare(NZ$)
4   (Direct)     NaN     NaN           133
7   (Direct)     NaN     NaN           163
```

```
8    (Direct)      NaN     NaN            163
10   (Direct)      NaN     NaN            193
11   (Direct)      NaN     NaN            193

       Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration   \
4       19/09/2019          AKL   9:00 AM          CHC  10:25 AM   1h 25m
7       19/09/2019          AKL  11:00 AM          CHC  12:25 PM   1h 25m
8       19/09/2019          AKL   7:00 PM          CHC   8:25 PM   1h 25m
10      19/09/2019          AKL  10:00 AM          CHC  11:25 AM   1h 25m
11      19/09/2019          AKL   4:00 PM          CHC   5:25 PM   1h 25m
…            …          …       …          …       …      …
162828  18/12/2019          ZQN   4:55 PM          WLG  10:10 PM   5h 15m
162829  18/12/2019          ZQN   9:35 AM          WLG   3:10 PM   5h 35m
162830  18/12/2019          ZQN  10:20 AM          WLG   6:10 PM   7h 50m
162831  18/12/2019          ZQN  10:20 AM          WLG   6:40 PM   8h 20m
162832  18/12/2019          ZQN   9:35 AM          WLG   6:10 PM   8h 35m

          Direct        Transit Baggage  Airfare(NZ$)
4       (Direct)            NaN     NaN           133
7       (Direct)            NaN     NaN           163
8       (Direct)            NaN     NaN           163
10      (Direct)            NaN     NaN           193
11      (Direct)            NaN     NaN           193
…          …            …     …           …
162828  (1 stop)  2h 15m in AKL     NaN           422
162829  (1 stop)  2h 35m in AKL     NaN           422
162830  (1 stop)  4h 50m in AKL     NaN           422
162831  (1 stop)  5h 20m in AKL     NaN           422
162832  (1 stop)  5h 35m in AKL     NaN           422

[157431 rows x 10 columns]
```

[4]:

```
df_sub = df_sub.drop('Transit',axis = 1)
df_sub = df_sub.drop('Baggage',axis = 1)
print(df_sub)
df_sub.isna().any()

df_sub = df_sub.dropna()

print(df_sub)
```

```
       Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration   \
4       19/09/2019          AKL   9:00 AM          CHC  10:25 AM   1h 25m
7       19/09/2019          AKL  11:00 AM          CHC  12:25 PM   1h 25m
8       19/09/2019          AKL   7:00 PM          CHC   8:25 PM   1h 25m
10      19/09/2019          AKL  10:00 AM          CHC  11:25 AM   1h 25m
11      19/09/2019          AKL   4:00 PM          CHC   5:25 PM   1h 25m
…            …          …       …          …       …      …
```

```
162828  18/12/2019          ZQN   4:55 PM          WLG  10:10 PM   5h 15m
162829  18/12/2019          ZQN   9:35 AM          WLG   3:10 PM   5h 35m
162830  18/12/2019          ZQN  10:20 AM          WLG   6:10 PM   7h 50m
162831  18/12/2019          ZQN  10:20 AM          WLG   6:40 PM   8h 20m
162832  18/12/2019          ZQN   9:35 AM          WLG   6:10 PM   8h 35m

          Direct  Airfare(NZ$)
4       (Direct)           133
7       (Direct)           163
8       (Direct)           163
10      (Direct)           193
11      (Direct)           193
…          …           …
162828  (1 stop)           422
162829  (1 stop)           422
162830  (1 stop)           422
162831  (1 stop)           422
162832  (1 stop)           422

[157431 rows x 8 columns]
```

```
[4]: Travel Date     False
     Dep. airport     True
     Dep. time       False
     Arr. airport     True
     Arr. time       False
     Duration        False
     Direct          False
     Airfare(NZ$)    False
     dtype: bool
```

```
       Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration  \
4       19/09/2019          AKL   9:00 AM          CHC  10:25 AM   1h 25m
7       19/09/2019          AKL  11:00 AM          CHC  12:25 PM   1h 25m
8       19/09/2019          AKL   7:00 PM          CHC   8:25 PM   1h 25m
10      19/09/2019          AKL  10:00 AM          CHC  11:25 AM   1h 25m
11      19/09/2019          AKL   4:00 PM          CHC   5:25 PM   1h 25m
…            …          …       …          …       …      …
162828  18/12/2019          ZQN   4:55 PM          WLG  10:10 PM   5h 15m
162829  18/12/2019          ZQN   9:35 AM          WLG   3:10 PM   5h 35m
162830  18/12/2019          ZQN  10:20 AM          WLG   6:10 PM   7h 50m
162831  18/12/2019          ZQN  10:20 AM          WLG   6:40 PM   8h 20m
162832  18/12/2019          ZQN   9:35 AM          WLG   6:10 PM   8h 35m

          Direct  Airfare(NZ$)
4       (Direct)           133
7       (Direct)           163
8       (Direct)           163
```

```
10      (Direct)            193
11      (Direct)            193
…          …           …
162828  (1 stop)            422
162829  (1 stop)            422
162830  (1 stop)            422
162831  (1 stop)            422
162832  (1 stop)            422

[157407 rows x 8 columns]
```

[5]:

```
#dataframe.drop(dataframe[dataframe['LND010200D'] == 0].index)
#dataframe.drop(dataframe[dataframe['population'] >= 1000000].index)
#subset_one = dataframe[dataframe['LND010200D'] != 0]
#subset_end = subset_one[subset_one['population'] < 1000000]
#subset_end.loc[:,'population-density'] = subset_end['population']/
 ↪subset_end['LND010200D']
#subset_end.loc[:,'case-ratio'] = subset_end['cases']/subset_end['population']
#subset_end = subset_end.assign(population_density=lambda x: x.population / x.
 ↪LND010200D)
#subset_end = subset_end.assign(case_ratio=lambda x: x.cases / x.population)

print(df_sub['Dep. airport'].unique())
print(df_sub['Arr. airport'].unique())
cols = df_sub.columns
print(cols)
print(df_sub.dtypes)

#print(df_sub.iloc[1]['Travel Date'].to_datetime.dt.day_name())
#print("2020-12-2".to_datetime.dt.day_name())
a = "19/9/2019"#"2020-10-10"
a = pd.to_datetime(a)
print(a.day_name())
```

```
['AKL' 'CHC' 'WLG' 'ZQN']
['CHC' 'DUD' 'NPE' 'NPL' 'NSN' 'PMR' 'WLG' 'ZQN' 'AKL']
Index(['Travel Date', 'Dep. airport', 'Dep. time', 'Arr. airport', 'Arr. time',
       'Duration', 'Direct', 'Airfare(NZ$)'],
      dtype='object')
Travel Date     object
Dep. airport    object
Dep. time       object
Arr. airport    object
Arr. time       object
Duration        object
Direct          object
Airfare(NZ$)     int64
dtype: object
```

Thursday

[6]:
```
df_sub ['Direct'] = df_sub['Direct'].replace(['(Direct)'],'0')
df_sub ['Direct'] = df_sub['Direct'].replace(['(1 stop)'],'1')
df_sub ['Direct'] = df_sub['Direct'].replace(['(2 stops)'],'2')
df_sub ['Direct'] = df_sub['Direct'].replace(['(3 stops)'],'3')

new_dep = []
for time in df_sub['Dep. time']:
    h,m = time.split(':')
    if time[len(time)-2:] == "AM":
        new_dep.append(h)
    elif time[len(time)-2:] == "PM":
        if h == '12':
            new_dep.append(h)
        else:
            new_dep.append(str(int(h)+12))
new_arr = []
for time in df_sub['Arr. time']:
    h,m = time.split(':')
    if time[len(time)-2:] == "AM":
        new_arr.append(h)
    elif time[len(time)-2:] == "PM":
        if h == '12':
            new_arr.append(h)
        else:
            new_arr.append(str(int(h)+12))
new_dur = []
for time in df_sub['Duration']:
    if 'h' in time:
        h,m = time.split(" ")
        mins = str(int(h[0:len(h)-1])*60+int(m[0:len(m)-1]))
        new_dur.append(mins)
    else:
        new_dur.append(time[0:len(time)-1])
df_sub['Duration'] = new_dur
df_sub['Dep. time'] = new_dep
df_sub['Arr. time'] = new_arr
#print(df_sub.iloc[1]['Travel Date'])
df_sub.head()
```

[6]:
```
   Travel Date Dep. airport Dep. time Arr. airport Arr. time Duration Direct  \
4   19/09/2019          AKL         9          CHC        10       85      0
7   19/09/2019          AKL        11          CHC        12       85      0
8   19/09/2019          AKL        19          CHC        20       85      0
10  19/09/2019          AKL        10          CHC        11       85      0
11  19/09/2019          AKL        16          CHC        17       85      0
```

```
    Airfare(NZ$)
4            133
7            163
8            163
10           193
11           193
```

[7]:

```
new_deptime = []
new_arrtime = []
new_date = []
for time in df_sub['Dep. time']:
    if int(time) <= 8:
        new_deptime.append("early")
    elif int(time) > 8 and int(time) < 12:
        new_deptime.append("morning")
    elif int(time) >= 12 and int(time) < 19:
        new_deptime.append("afternoon")
    else:
        new_deptime.append("late")

for time in df_sub['Arr. time']:
    if int(time) <= 8:
        new_arrtime.append("early")
    elif int(time) > 8 and int(time) < 12:
        new_arrtime.append("morning")
    elif int(time) >= 12 and int(time) < 19:
        new_arrtime.append("afternoon")
    else:
        new_arrtime.append("late")

for time in df_sub['Travel Date']:
    time = pd.to_datetime(time, infer_datetime_format=True)
    time = time.day_name()
    new_date.append(time)

df_sub['Dep. time'] = new_deptime
df_sub['Arr. time'] = new_arrtime

df_sub['Travel Date'] = new_date
df_sub.head()
```

[7]:

```
   Travel Date Dep. airport  Dep. time Arr. airport  Arr. time Duration  \
4     Thursday          AKL    morning          CHC    morning       85
7     Thursday          AKL    morning          CHC  afternoon       85
8     Thursday          AKL       late          CHC       late       85
10    Thursday          AKL    morning          CHC    morning       85
```

```
11     Thursday           AKL   afternoon          CHC   afternoon       85

   Direct  Airfare(NZ$)
4        0           133
7        0           163
8        0           163
10       0           193
11       0           193
```

[8]:

```
X_train, X_test, y_train, y_test = train_test_split(df_sub.drop(columns =␣
↪['Airfare(NZ$)','Travel Date','Dep. airport','Dep. time','Arr. airport','Arr.
↪ time'],axis = 1), df_sub['Airfare(NZ$)'], test_size=0.25, random_state = 2)
X_train
X_test
y_train
y_test
```

[8]:

```
        Duration  Direct
84951        270       1
115341       275       1
54025        540       1
22514        485       1
141047       370       1
…            …     …
34977        435       1
87284        190       1
99084        210       1
104309       305       1
92302        165       1

[118055 rows x 2 columns]
```

[8]:

```
        Duration  Direct
105525       270       1
141588       695       1
68115        270       2
137376       995       1
161697      1375       2
…            …     …
122566       195       1
62570        485       2
101606       815       1
34487        920       1
58224        220       1

[39352 rows x 2 columns]
```

```
[8]: 84951     517
     115341    391
     54025     512
     22514     338
     141047    391
              …
     34977     530
     87284     272
     99084     599
     104309    370
     92302     391
     Name: Airfare(NZ$), Length: 118055, dtype: int64
```

```
[8]: 105525    391
     141588    412
     68115     662
     137376    352
     161697    402
              …
     122566    502
     62570     293
     101606    433
     34487     492
     58224     462
     Name: Airfare(NZ$), Length: 39352, dtype: int64
```

```
[9]: model = LinearRegression(fit_intercept = True)

     model.fit(X_train, y_train)

     model.score(X_train, y_train)

     model.coef_ # this is beta 1, the slope of the regression function

     model.intercept_ # this is beta 0
```

```
[9]: LinearRegression()

[9]: 0.2760132742657364

[9]: array([-1.03556996e-01,  1.96533966e+02])

[9]: 260.70910048597125
```

```
[10]: test_output = pd.DataFrame(model.predict(X_test), index = X_test.index, columns␣
        ↪= ['predict Airface in NZ$'])
      test_output.head()
```

[10]:

```
        predict Airface in NZ$
105525                  429.28
141588                  385.27
68115                   625.82
137376                  354.20
161697                  511.39
```

[11]:

```
test_output = test_output.merge(y_test, left_index = True, right_index = True)
test_output.head()
mean_absolute_error = abs(test_output['predict Airface in NZ$'] -␣
 ↪test_output['Airfare(NZ$)']).mean()
print('Mean absolute error is ')
print(mean_absolute_error)
print('Fraction MAE is ')
print(mean_absolute_error / test_output['Airfare(NZ$)'].mean())
```

[11]:

```
        predict Airface in NZ$  Airfare(NZ$)
105525                  429.28           391
141588                  385.27           412
68115                   625.82           662
137376                  354.20           352
161697                  511.39           402
```

```
Mean absolute error is
108.68246539502618
Fraction MAE is
0.25905943887789995
```

[12]:

```
# define function to import viz libraries
import plotly
plotly.offline.init_notebook_mode(connected=True)
from plotly.graph_objs import *
from plotly import tools
import plotly.graph_objects as go
import seaborn as sns
```

[13]:

```
df_sub.head()
from sklearn.preprocessing import OneHotEncoder

def get_ohe(df, col):
    ohe = OneHotEncoder(drop='first', handle_unknown='error', sparse=False,␣
 ↪dtype='int')
    ohe.fit(df[[col]])
    temp_df = pd.DataFrame(data=ohe.transform(df[[col]]), columns=ohe.
 ↪get_feature_names())
    # If you have a newer version, replace with columns=ohe.
 ↪get_feature_names_out()
```

```
    df.drop(columns=[col], axis=1, inplace=True)
    df = pd.concat([df.reset_index(drop=True), temp_df], axis=1)
    return df
```

```
[13]:    Travel Date Dep. airport  Dep. time Arr. airport  Arr. time Duration  \
      4     Thursday          AKL    morning          CHC    morning       85
      7     Thursday          AKL    morning          CHC  afternoon       85
      8     Thursday          AKL       late          CHC       late       85
      10    Thursday          AKL    morning          CHC    morning       85
      11    Thursday          AKL  afternoon          CHC  afternoon       85

         Direct  Airfare(NZ$)
      4       0           133
      7       0           163
      8       0           163
      10      0           193
      11      0           193
```

```
[14]: df_sub = get_ohe(df_sub, 'Travel Date')
      df_sub = get_ohe(df_sub, 'Dep. airport')
      df_sub = get_ohe(df_sub, 'Dep. time')
      df_sub = get_ohe(df_sub, 'Arr. airport')
      df_sub = get_ohe(df_sub, 'Arr. time')

      df_sub.head(5)


      X_train, X_test, y_train, y_test = train_test_split(df_sub.drop(columns =␣
        ↪['Airfare(NZ$)']),df_sub['Airfare(NZ$)'], test_size=0.25,random_state = 2)
      X_train
      X_test
      y_train
      y_test
```

```
/Users/shanghuahao/opt/anaconda3/lib/python3.9/site-
packages/sklearn/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0
and will be removed in 1.2. Please use get_feature_names_out instead.

/Users/shanghuahao/opt/anaconda3/lib/python3.9/site-
packages/sklearn/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0
and will be removed in 1.2. Please use get_feature_names_out instead.

/Users/shanghuahao/opt/anaconda3/lib/python3.9/site-
```

```
packages/sklearn/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0
and will be removed in 1.2. Please use get_feature_names_out instead.

/Users/shanghuahao/opt/anaconda3/lib/python3.9/site-
packages/sklearn/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0
and will be removed in 1.2. Please use get_feature_names_out instead.

/Users/shanghuahao/opt/anaconda3/lib/python3.9/site-
packages/sklearn/utils/deprecation.py:87: FutureWarning:

Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0
and will be removed in 1.2. Please use get_feature_names_out instead.
```

```
[14]:   Duration Direct  Airfare(NZ$)  x0_Monday  x0_Saturday  x0_Sunday  \
      0        85      0           133          0            0          0
      1        85      0           163          0            0          0
      2        85      0           163          0            0          0
      3        85      0           193          0            0          0
      4        85      0           193          0            0          0

         x0_Thursday  x0_Tuesday  x0_Wednesday  x0_CHC  …  x0_DUD  x0_NPE  x0_NPL  \
      0            1           0             0       0  …       0       0       0
      1            1           0             0       0  …       0       0       0
      2            1           0             0       0  …       0       0       0
      3            1           0             0       0  …       0       0       0
      4            1           0             0       0  …       0       0       0

         x0_NSN  x0_PMR  x0_WLG  x0_ZQN  x0_early  x0_late  x0_morning
      0       0       0       0       0         0        0           1
      1       0       0       0       0         0        0           0
      2       0       0       0       0         0        1           0
      3       0       0       0       0         0        0           1
      4       0       0       0       0         0        0           0

       [5 rows x 26 columns]
```

```
[14]:          Duration Direct  x0_Monday  x0_Saturday  x0_Sunday  x0_Thursday  \
      82184         270      1          0            0          0            0
      111533        275      1          0            0          1            0
      52300         540      1          0            0          0            0
      21796         485      1          1            0          0            0
      136356        370      1          0            1          0            0
```

```
…         …    …        …          …        …          …
33867       435      1          1            0          0            0
84434       190      1          1            0          0            0
95816       210      1          0            0          1            0
100879      305      1          1            0          0            0
89256       165      1          0            1          0            0

        x0_Tuesday  x0_Wednesday  x0_CHC  x0_WLG   … x0_DUD  x0_NPE  x0_NPL   \
82184            0             0       0       0   …      0       0       0
111533           0             0       0       0   …      1       0       0
52300            0             1       0       0   …      0       1       0
21796            0             0       1       0   …      0       1       0
136356           0             0       1       0   …      0       0       0
…            …           …     …     …  …     …     …     …
33867            0             0       1       0   …      0       0       0
84434            0             0       1       0   …      0       0       0
95816            0             0       0       1   …      0       0       0
100879           0             0       1       0   …      0       0       0
89256            0             0       1       0   …      0       0       0

        x0_NSN  x0_PMR  x0_WLG  x0_ZQN  x0_early  x0_late  x0_morning
82184        0       0       0       0         0        0           0
111533       0       0       0       0         0        1           0
52300        0       0       0       0         0        0           0
21796        0       0       0       0         0        0           0
136356       0       0       0       0         0        0           0
…        …     …     …     …       …      …         …
33867        0       0       0       1         0        1           0
84434        0       0       1       0         0        0           0
95816        0       0       0       0         0        0           0
100879       0       1       0       0         0        0           0
89256        0       0       0       0         0        0           0

[118055 rows x 25 columns]
```

[14]:

```
       Duration Direct  x0_Monday  x0_Saturday  x0_Sunday  x0_Thursday   \
102055      270      1          0            0          0            0
136881      695      1          0            1          0            0
65915       270      2          0            0          0            0
132830      995      1          0            0          0            1
156307     1375      2          0            0          0            0
…         …    …        …          …        …          …
118508      195      1          0            0          0            0
60542       485      2          1            0          0            0
98254       815      1          0            0          0            0
33391       920      1          1            0          0            0
56343       220      1          0            0          0            0
```

```
        x0_Tuesday  x0_Wednesday  x0_CHC  x0_WLG  …  x0_DUD  x0_NPE  x0_NPL  \
102055           0             1       0       0  …       0       0       0
136881           0             0       0       1  …       0       1       0
65915            0             1       0       0  …       0       1       0
132830           0             0       0       0  …       0       0       1
156307           1             0       0       0  …       1       0       0
…              …           …     …     …  …   …     …    …
118508           0             1       0       0  …       0       0       0
60542            0             0       0       0  …       1       0       0
98254            0             0       0       0  …       0       0       0
33391            0             0       1       0  …       1       0       0
56343            0             0       0       1  …       0       0       0

        x0_NSN  x0_PMR  x0_WLG  x0_ZQN  x0_early  x0_late  x0_morning
102055       0       0       0       0         0        0           0
136881       0       0       0       0         0        0           0
65915        0       0       0       0         0        0           0
132830       0       0       0       0         1        0           0
156307       0       0       0       0         0        0           0
…          …     …     …     …       …      …        …
118508       0       0       1       0         0        0           0
60542        0       0       0       0         0        1           0
98254        0       0       0       1         0        1           0
33391        0       0       0       0         0        0           1
56343        0       0       0       0         0        0           1

[39352 rows x 25 columns]
```

```
[14]: 82184     517
      111533    391
      52300     512
      21796     338
      136356    391
               …
      33867     530
      84434     272
      95816     599
      100879    370
      89256     391
      Name: Airfare(NZ$), Length: 118055, dtype: int64
```

```
[14]: 102055    391
      136881    412
      65915     662
      132830    352
      156307    402
```

```
          …
118508    502
60542     293
98254     433
33391     492
56343     462
Name: Airfare(NZ$), Length: 39352, dtype: int64
```

[15]:
```
model = LinearRegression(fit_intercept = True)

model.fit(X_train, y_train)

model.score(X_train, y_train)

model.coef_ # this is beta 1, the slope of the regression function

model.intercept_ # this is beta 0
```

[15]:
```
LinearRegression()
```

[15]:
```
0.369857806426307944
```

[15]:
```
array([-9.28523597e-02,  2.11655350e+02, -3.56602260e+01, -2.66531262e+01,
       -1.07045402e+01, -4.21580009e+01, -7.12800966e+01, -5.38788289e+01,
        3.87622290e+00, -3.26226022e+01, -6.04674975e+00, -4.21339308e+01,
       -9.62636142e+00, -2.00377868e+01,  9.38429099e+00, -3.40049900e+01,
        3.66039527e+01, -6.80418342e+01, -7.62126191e+01, -4.09387937e+01,
        7.36553468e+00, -1.41276418e+01,  8.85913291e+00, -1.95351202e+01,
        2.88913261e+01])
```

[15]:
```
329.34194908948507
```

[16]:
```
test_output = pd.DataFrame(model.predict(X_test), index = X_test.index, columns␣
 ↪= ['predict Airface in NZ$'])
test_output.head()
```

[16]:
```
        predict Airface in NZ$
102055                  471.43
136881                  411.66
65915                   662.13
132830                  341.22
156307                  513.65
```

[17]:
```
test_output = test_output.merge(y_test, left_index = True, right_index = True)
test_output.head()
mean_absolute_error = abs(test_output['predict Airface in NZ$'] -␣
 ↪test_output['Airfare(NZ$)']).mean()
```

```python
print('Mean absolute error is ')
print(mean_absolute_error)
print('Fraction MAE is ')
print(mean_absolute_error / test_output['Airfare(NZ$)'].mean())
```

[17]:

```
        predict Airface in NZ$  Airfare(NZ$)
102055                  471.43           391
136881                  411.66           412
65915                   662.13           662
132830                  341.22           352
156307                  513.65           402

Mean absolute error is
101.16054014322361
Fraction MAE is
0.24112990693426173
```

[ ]:

```

```

[ ]:

```

```